Best-of-Evidence: Best-of-N Selection under Partial Verification

Cenwei Zhang[1†] **Teng Fang**[1] **Yuxia Wang**[2] **Derek Li**[1] **Bryan Dai**[1‡] **Lei You**[3‡]

[1]IQuest Research [2]INSAIT [3]Technical University of Denmark

cwzhang2001@gmail.com yuxia.wang@insait.ai

{tfang,jdli,cbdai}@iquestlab.com leiyo@dtu.dk

[†]Work done during an internship at IQuest Research. [‡]Corresponding authors.

Abstract

BoN improves model outputs by sampling several candidates and selecting one with a proxy score, but it assumes that complete candidates can be evaluated reliably. Many vision-language tasks instead provide only partial verification: a finding, span, value, region, or relation may be checkable even when no dependable whole-response verifier exists. Moreover, the same claim may recur across candidates with opposing stances, allowing one observation to support part of the pool and contradict another. We introduce *Best-of-Evidence* (BoE), an inference-time selection framework that keeps the BoN candidate pool fixed, represents reusable claims with a signed candidate–factor graph, and allocates a limited budget to evidence actions that can change the final choice. BoE formalizes selection under partial verification and provides a practical score-based controller, with the zero-budget case recovering the underlying BoN decision. Theoretically, we show that residual evidence capacity limits any evidence-driven improvement and that shared factor queries can achieve an $O(\log K)$ versus $\Theta(K)$ query separation in a factor-code model. Common-ledger experiments on four medical VQA settings show that BoE can improve fixed-pool selection and rescue some BoN failures when evidence is reliable, contrastive, and decision-relevant, while also revealing the channel-quality and candidate-generation limits that prevent universal gains.

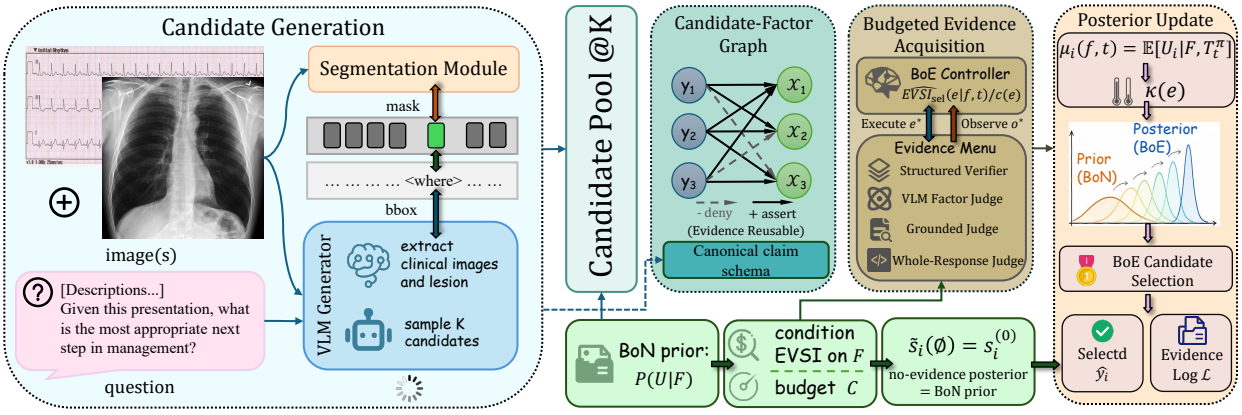

Figure 1: **Overview of Best-of-Evidence under partial verification.** From left to right, a VLM generates a fixed candidate pool, with segmentation used only to prepare grounded views. The cheap BoN context initializes the ranking. Candidate claims are merged into a signed candidate–factor graph, allowing one observation to affect multiple candidates in different directions. Under a budget, BoE selects among heterogeneous checks, updates the ranking with the acquired evidence, and returns the selected candidate together with an evidence log.

1 Introduction

Language models often produce several plausible solutions even when a single decode is unreliable. When changing or retraining the generator is costly, sampling multiple attempts and selecting among them offers a simple way to trade test-time compute for output quality [1, 2, 3, 4]. *Best-of-N* (BoN) follows this pattern: it samples K candidates, scores each with a proxy, and returns the highest-scoring one. The proxy may be a majority vote, reward model, verifier, process reward model, or another judge [5, 6, 7]. BoN is attractive because it leaves the generator fixed and spends additional compute only on test-time generation and selection.

The limitation is the unit of verification. Standard outcome-level BoN assumes that each complete candidate admits a reliable scalar score [5]. This is natural for code, arithmetic, and some closed-form tasks [2, 8, 9], but less natural for many vision-language model (VLM) tasks. A candidate may be partly right and partly wrong [3, 7, 10, 11, 12]. A report may contain a correct modality claim and an unsupported finding [13, 14, 15]; a chart explanation may read one value correctly but compare two values incorrectly [16, 17]; and a document answer may depend on multiple OCR spans and a layout relation [18, 19]. In such cases, no cheap whole-response verdict may exist, although a crop, lookup, region, span, or isolated claim can still be checked [11, 20, 21, 22]. Selection therefore faces a granularity mismatch: candidates are complete objects, while verification arrives through costed local views.

Best-of-Evidence (BoE) addresses this mismatch while keeping the sampled pool fixed. The cheap BoN context initializes candidate ranking; BoE then represents canonical claims and candidate stances, and spends a limited budget on checks that may change the final choice. In an ideal Bayesian formulation, observations update posterior mean utilities. The practical controller approximates this objective with discrimination-weighted scores and a plug-in value-of-information rule. With zero evidence budget, BoE recovers the underlying BoN decision.

This view builds on recent reward-evaluation systems that organize heterogeneous tools, references, checklists, and verifiers into grounded evaluation procedures [23, 24, 25]. These systems show how evidence can be gathered and aggregated; the remaining question here is which checks are worth acquiring when no dependable whole-response score is available. A lookup, local judge, or whole-response evaluator is useful only insofar as its possible outcomes can alter selection.

BoE represents that selection structure with a signed candidate–factor graph. Factor nodes are inspectable claims, and edges record whether each candidate asserts, denies, or is unrelated to a claim [26]. One observation can therefore support part of the pool and contradict another. Useful reuse is not determined by graph degree alone: the checked claim must separate candidates that remain competitive under the cheap prior. This signed sharing permits compressive verification [27] without assuming that every candidate pool forms an exact latent factor code.

We make three contributions in this paper:

- We formulate selection under partial verification and introduce a budgeted score-based controller that reuses signed local evidence across a fixed BoN candidate pool.
- We establish a residual-information upper bound, a scoped factor-code query separation, and a local account of decision-effective evidence reuse.
- We evaluate BoE with common evidence ledgers on four medical VQA settings, characterizing fixed-ledger policy differences and the channel and candidate-pool limits of partial verification.

2 Selection under Partial Verification

Before introducing BoE, we first define the partial-verification selection problem it approximates. We proceed from the fixed candidate pool, to the cheap prior, to shared verifiable claims, and finally to the budgeted selection objective that Section 3 seeks to approximate.

Notation conventions. Random objects are uppercase and their realizations are lowercase when the distinction matters. We use K for the candidate-pool size, U_i for latent candidate utility, μ_i for an exact Bayesian posterior mean, and $\widetilde{s}_i$ for the uncalibrated score used in ledger replay. The symbol π denotes an evidence-acquisition policy, whereas p_ϕ^{gen} denotes the candidate generator. Candidate, claim, acquisition-round, dataset, and question indices are i, j, t, d, r; m denotes a terminal selection size and q a query count. We use E_t and O_t for random acquired actions and outcomes, e and o for their realizations, $\mathtt{t}$ for a realized transcript, and O_e for the prospective outcome of acquiring action e next. Unless explicitly stated in bits, mutual information is measured in nats.

2.1 Candidate pool and cheap prior

We first fix the objects among which selection is performed. Let (Ω, X) denote a random task instance. The record X contains the multimodal input, question, and admissible same-instance metadata, but not the benchmark reference. Conditional on $X = x \in \mathcal{X}$, the generator samples the joint candidate pool

$$\mathbf{Y}_{1:K} \mid X = x \sim p_\phi^{\text{gen}}(\cdot \mid x). \tag{1}$$

This equation fixes the pool used by all later verification and selection. A realized pool $\mathbf{y}_{1:K}$ remains unchanged throughout, and no conditional independence among its candidates is required.

A candidate may be a short answer, report, box set, mask, or another structured output. The latent state Ω contains the facts or reference information needed to define correctness, but neither the selector nor an evidence action may query it. Candidate utility is

$$U_i = u(\Omega, \mathbf{Y}_i) \in [0, 1], \qquad \mathbf{U} = (U_1, \ldots, U_K). \tag{2}$$

This defines the hidden target that evidence and selection seek to infer.

Before purchasing evidence, the selector observes only $F = f_0(X, \mathbf{Y}_{1:K})$. The fixed cheap interface f_0 may expose candidate outputs, answer frequencies, validity flags, self-consistency statistics, and other declared zero-cost features, but not the complete record X. It induces

$$P(\mathbf{U} \mid F), \qquad \mu_i^{\text{prior}}(F) = \mathbb{E}[U_i \mid F]. \tag{3}$$

This is the zero-evidence Bayesian ranking. The problem is nontrivial only when $\mathbf{U}$ is not almost surely determined by F. An ideal Bayesian BoN selector maximizes $\mu_i^{\text{prior}}(F)$ with fixed tie-breaking. The ledger implementation instead initializes an answer-frequency score $s_i^{(0)}$ whose maximizer agrees with raw majority; it is not assumed that $s_i^{(0)} = \mu_i^{\text{prior}}(F)$.

2.2 Signed shared claims

We next define what partial verification can address. Let $\chi_1, \ldots, \chi_M$ be canonical verifiable propositions about Ω, called factors when represented as graph nodes, and let $\mathbf{\Theta} = (\Theta_1, \ldots, \Theta_M) \in \{0, 1\}^M$ denote their latent truth values. Examples include a chart value, OCR span, spatial relation, modality, anatomical site, or visual finding.

The signed incidence matrix

$$\mathbf{B} \in \{-1, 0, +1\}^{K \times M} \tag{4}$$

records candidate stance: $B_{ij} = +1$ if candidate i asserts χ_j, $B_{ij} = -1$ if it denies χ_j, and $B_{ij} = 0$ if the claim is irrelevant. The same information is represented as the signed candidate–factor graph

$$G_{\text{cf}} = (\mathcal{V}_c, \mathcal{V}_f, \mathcal{E}_{\text{cf}}), \tag{5}$$

where $\mathcal{V}_c$ contains candidate nodes, $\mathcal{V}_f$ contains claim nodes, and signed edges correspond to nonzero entries of $\mathbf{B}$.

Together, these equations determine how one observation is reused. Confirming claim j supports candidates with $B_{ij} = +1$, contradicts those with $B_{ij} = -1$, and leaves those with $B_{ij} = 0$ unchanged;

rejection reverses the two nonzero directions. The observation is acquired once and applied through all incident edges.

The graph is an evidence interface, not a complete causal or generative model of utility. It need not explain every component of U_i or require $\boldsymbol{\Theta}$ to determine $\mathbf{U}$. Sharing alone is also insufficient: if all competitive candidates take the same stance, a check may move their scores together without changing the winner. Useful shared evidence must therefore be residual and contrastive under the cheap prior.

2.3 Costed evidence and the decision objective

We finally define how additional information is acquired. Let $\mathcal{A}$ denote the evidence-action menu. At round t, let E_t be the random action, O_t its random outcome, and $T^{\pi}_{t-1} = (E_1, O_1, \ldots, E_{t-1}, O_{t-1})$ the purchased transcript prefix. The executor constructs an action-specific view of the same record and returns an observation:

$$Z_t = \mathrm{view}_{E_t}(X, \mathbf{Y}_{1:K}, T^{\pi}_{t-1}), \qquad O_t = \psi_{E_t}(Z_t, F, T^{\pi}_{t-1}; \mathsf{Seed}^{\mathrm{exec}}_t). \tag{6}$$

The equation separates what an action is allowed to inspect from the verdict it returns. A view may be a crop, mask-guided region, metadata field, structured lookup, or human inspection of the same X. Neither map may query Ω, $\mathbf{U}$, or the benchmark reference. Factor actions inspect claim nodes, whereas candidate actions inspect complete responses. Localization only determines the view; it is not an additional observation.

An access-admissible policy observes only F, purchased action–outcome pairs, and independent private randomization. It produces a terminal transcript T_π with realized cost at most C. The complete measurability, stopping-time, and transcript definitions are given in Appendix A.1; let Π^{acc}_C denote this policy class.

After observing T_π, the Bayes action selects the candidate with the largest conditional mean utility. The ideal budgeted value is

$$\mathrm{OPT}^{\mathrm{sel}}_C = \sup_{\pi \in \Pi^{\mathrm{acc}}_C} \underbrace{\mathbb{E}\Big[\max_{1 \le i \le K} \mathbb{E}[U_i \mid F, T_\pi]\Big]}_{\substack{\text{expected utility of the candidate selected} \\ \text{after the purchased evidence}}} . \tag{7}$$

Equation (7) is the key to Section 3. For a fixed evidence policy, the inner conditional expectation gives the posterior mean utility of each candidate after the purchased observations, and the maximum selects the best candidate under that updated belief. The outer expectation averages over task instances, candidate pools, evidence outcomes, and policy randomness. The supremum asks which budget-feasible evidence policy gives the best final selection on average.

With $C = 0$, the transcript is empty and the objective reduces to ideal Bayesian BoN. Exact optimization is generally intractable because actions have heterogeneous costs, correlated effects, and adaptive availability. Section 3 therefore uses a sequential value-of-information approximation.

3 Best-of-Evidence

BoE is a practical sequential approximation to Equation (7). As illustrated in Figure 1, it builds signed shared claims from a fixed candidate pool, attaches heterogeneous checks, allocates a budget according to their predicted effect on selection, and re-ranks the unchanged pool after each observation.

3.1 Bayesian evidence allocation

BoE canonicalizes equivalent claims, preserves candidate stance, and associates each claim with admissible evidence actions. The menu may combine structured verification, whole-image or grounded VLM

judgments, and whole-response judging. An unresolved or inapplicable claim produces no applicable signal rather than a contradiction.

For a realized cheap context $F = f$ and transcript $\mathtt{t} = ((e_1, o_1), \ldots, (e_t, o_t))$, the ideal posterior mean utility is

$$\mu_i(f, \mathtt{t}) = \mathbb{E}[U_i \mid F = f, T_t^\pi = \mathtt{t}]. \tag{8}$$

Let $V_{\mathrm{sel}}(f, \mathtt{t}) = \max_i \mu_i(f, \mathtt{t})$ denote the value of selecting immediately. For a feasible unrevealed action e, its exact expected value of sample information is

$$\mathrm{EVSI}_{\mathrm{sel}}(e \mid f, \mathtt{t}) = \mathbb{E}_{O_e \mid F=f, T_t^\pi = \mathtt{t}}[V_{\mathrm{sel}}(f, \mathtt{t} \oplus (e, O_e))] - V_{\mathrm{sel}}(f, \mathtt{t}). \tag{9}$$

This quantity measures the expected change in final selection value rather than claim accuracy or graph degree alone. BoE applies this principle sequentially: it acquires a feasible action with high estimated value per cost, updates the transcript and candidate scores, and stops when the remaining budget or estimated value no longer justifies another check. The formal greedy rule, stopping condition, and complete replay procedure are given in Appendix A.2 and Algorithm 1.

3.2 Discrimination-weighted ledger replay

The benchmark replay does not fit a complete observation model for $P(\Omega, \mathbf{U}, T_\pi \mid F)$. It initializes $\widetilde{s}_i(\varnothing) = s_i^{(0)}$ and maps each outcome $o \in \mathcal{O}_e$ to a signed candidate effect $\alpha_{ie}(o) \in \{-1, 0, +1\}$. Each action also inherits a pooled outcome mass $\widehat{p}_e$ and a selection-side discrimination index $\kappa(e) \in (0, 1)$. Writing $\mathrm{wt}(e) = \mathrm{logit}\, \kappa(e)$, the implemented update is

$$\mathrm{logit}\, \widetilde{s}_i(\mathtt{t} \oplus (e, o)) = \mathrm{logit}\, \widetilde{s}_i(\mathtt{t}) + \alpha_{ie}(o)\, \mathrm{wt}(e). \tag{10}$$

Thus $\alpha_{ie}(o) = 0$ or $\kappa(e) = 0.5$ gives no update, while $\kappa(e) < 0.5$ reverses an anti-correlated channel. The controller evaluates all possible outcomes with the following plug-in score.

Implemented replay action value

$$\widehat{\mathrm{EVSI}}_{\mathrm{sel}}(e \mid f, \mathtt{t}) = \underbrace{\sum_{o \in \mathcal{O}_e} \widehat{p}_e(o) \max_i \widetilde{s}_i(\mathtt{t} \oplus (e, o))}_{\substack{\text{expected best score} \\ \text{after check } e}} - \underbrace{\max_i \widetilde{s}_i(\mathtt{t})}_{\text{current best score}}. \tag{11}$$

Feasible actions are ranked by $\widehat{\mathrm{EVSI}}_{\mathrm{sel}}(e \mid f, \mathtt{t})/c(e)$. The hat marks a ledger-instantiated score approximation rather than exact Bayesian EVSI.

The pooled $\widehat{p}_e$ is not a history-conditioned predictive model, and $\kappa(e)$ measures candidate discrimination rather than factor-truth accuracy or a likelihood parameter. Accordingly, $\widetilde{s}_i$ is an evidence-updated selection score, not a calibrated posterior probability. Appendix A.2 gives the full score construction and its relation to the exact Bayesian quantities. The evidence log records each acquired action, source, outcome, cost, discrimination index, and affected candidates. It contains no chain-of-thought text.

4 When Can Shared Evidence Help?

Shared evidence can help only if it is decision-relevant, structurally reusable, and informative beyond the cheap context. We summarize these three mechanisms and defer their formal statements and proofs to Appendix A.

4.1 Decision-effective signed reuse

At an active history $h = (f, \mathtt{t})$, let $\mathbf{b}_j$ be column j of the signed candidate–claim matrix. A positive semidefinite matrix $\mathbf{W}_h$ weights score directions by their effect on the current selection. We define

$$R_{\mathrm{eff}}(j; h) = \mathbf{b}_j^\top \mathbf{W}_h \mathbf{b}_j. \tag{12}$$

Thus, a claim can touch many candidates yet have little value if it moves them in a common direction.

When is a shared claim worth checking?

For a factor action e targeting claim $j(e)$, let $\iota_e(h)$ be the conditional variance of its centered score innovation and $c(e)$ its cost. The local model yields, up to a common factor $1/2$,

$$D_{\text{factor}}(e;h) = \frac{\underbrace{\iota_e(h)}_{\substack{\text{unpredictable}\\ \text{score signal}}} \quad \underbrace{R_{\text{eff}}(j(e);h)}_{\substack{\text{signed decision}\\ \text{contrast}}}}{\underbrace{c(e)}_{\text{cost}}}. \tag{13}$$

This is a local explanatory proxy, not a Taylor expansion or the implemented controller objective. The replay controller does not estimate $\mathbf{W}_h$, R_{eff}, or $\iota_e(h)$.

The formula explains why useful reuse requires signal, signed contrast, and favorable cost rather than graph degree alone. Appendix A compares this density with candidate-level checks.

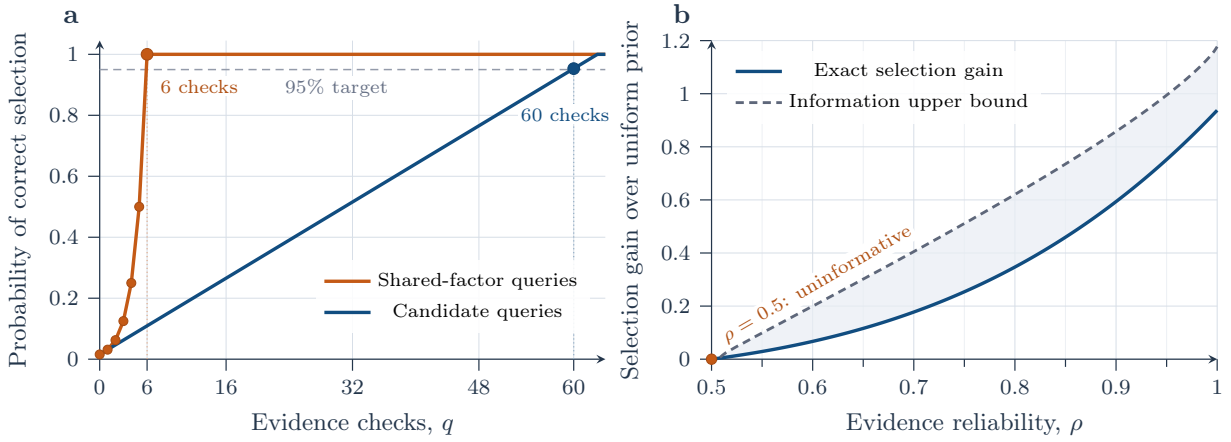

Figure 2: **Exact consequences of the finite constructions. a**, Six shared-factor queries recover a six-bit code; candidate queries need 60 checks to exceed 95% success. **b**, In the four-factor noisy model, the exact gain $\rho^4 - 1/16$ stays below the information upper bound and vanishes with an uninformative channel.

4.2 A scoped query-complexity separation

We prove a structural witness in which $K = 2^M$ candidates encode all assignments of M binary claims. Exact factor queries reveal shared coordinates, whereas exact candidate queries test complete assignments.

Exact takeaway. For target success at least $1 - \varepsilon$, $\varepsilon \in [0, 1)$,

$$\underbrace{Q_{\text{factor}} = \log_2 K}_{\text{shared-coordinate checks}} \qquad \text{whereas} \qquad \underbrace{Q_{\text{candidate}} \geq \lceil (1-\varepsilon)K \rceil - 1}_{\text{whole-candidate checks}}.$$

This proves a possible $O(\log K)$ versus $\Theta(K)$ gap, not that arbitrary VLM pools form complete codes. Figure 2 shows this construction and a separate noisy-factor model; the closed-form calculations and access-admissible proof appear in Appendix A. The noisy-model parameter ρ is distinct from the empirical replay index κ.

4.3 Residual-information limit

For an adaptive policy π, let T_π contain its acquired actions and observations. The residual evidence capacity under budget C is

$$\Lambda_C = \sup_{\pi \in \Pi_C^{\mathrm{acc}}} \mathrm{I}(\mathbf{U}; T_\pi \mid F). \tag{14}$$

It removes information already present in F and accounts for redundancy among adaptive observations.

Under the conditional sub-Gaussian assumption in Appendix A, we prove the following evidence-only ceiling for a terminal selection of size m.

The information ceiling

$$\mathbb{E}[V_m^\pi(F, T_\pi) - V_m(F)] \le \underbrace{\sqrt{\frac{m}{2}\,\mathrm{I}(\mathbf{U}; T_\pi \mid F)}}_{\substack{\text{from acquired information} \\ \text{beyond } F}} \le \underbrace{\sqrt{\frac{m}{2}\Lambda_C}}_{\substack{\text{from best accessible evidence} \\ \text{within budget}}}. \tag{15}$$

Small residual information rules out a large evidence-driven improvement.

We also prove a total-throughput upper law combining cheap-context information and Λ_C. These are necessary information constraints, not achievability guarantees or controller objectives; all assumptions and proofs are given in Appendix A.

5 Experiments

We evaluate whether budgeted evidence improves selection after the candidate pool and all potential observations are fixed. This common-ledger design isolates the selection mechanism rather than end-to-end variation from new candidate samples. It also probes the account in Section 4: useful gains require residual, decision-relevant evidence and a correct candidate that selection can recover. Detailed protocols, the historical SLAKE pilot, and other information appear in Appendix B.

5.1 Benchmark design

Evaluation cells and models. The evaluation covers VQA-Med, PathVQA, an option-hidden PMC-VQA protocol, and MedXpertQA-MM [28, 29, 30, 31]. PMC-VQA hides the original options and uses the correct option text as the open-answer reference; MedXpertQA-MM retains five options and one to six images. All four cells use Qwen3-VL-30B-A3B as the candidate generator and Qwen3-VL-235B-A22B as the evidence judge. We sample $K = 16$ candidates at temperature 1.1 and replay budgets $C \in \{1, 2, 4, 8, 16\}$, with $C = 16$ as the main comparison. Table 1 summarizes the resulting candidate pools and their selection ceilings.

Table 1: **Main evaluation cells and candidate-generation health.** Parsed is the fraction of generations with a recovered structured output; Answer is the fraction with a usable final answer. Oracle@K is the fraction of questions whose candidate pool contains at least one correct answer.

Dataset	Protocol	Generator / judge	n_d	Parsed (%)	Answer (%)	Oracle@K (%)
VQA-Med	Open answer, fixed battery	30B / 235B	2,334	95.1	97.7	82.1
PathVQA	Open answer, fixed battery	30B / 235B	9,903	94.0	99.1	61.5
PMC-VQA	Option-hidden open protocol	30B / 235B	10,000	88.2	98.5	65.1
MedXpertQA-MM	Single-answer five-way MCQ, 1–6 images	30B / 235B	2,000	96.2	94.5	64.7

Ledger, evidence, and policies. For each question, we cache one candidate pool, its signed factor graph, and all available evidence outcomes, channel identities, costs, and evaluation utilities. Replay policies therefore share the same generation and potential observations; they differ only in which entries

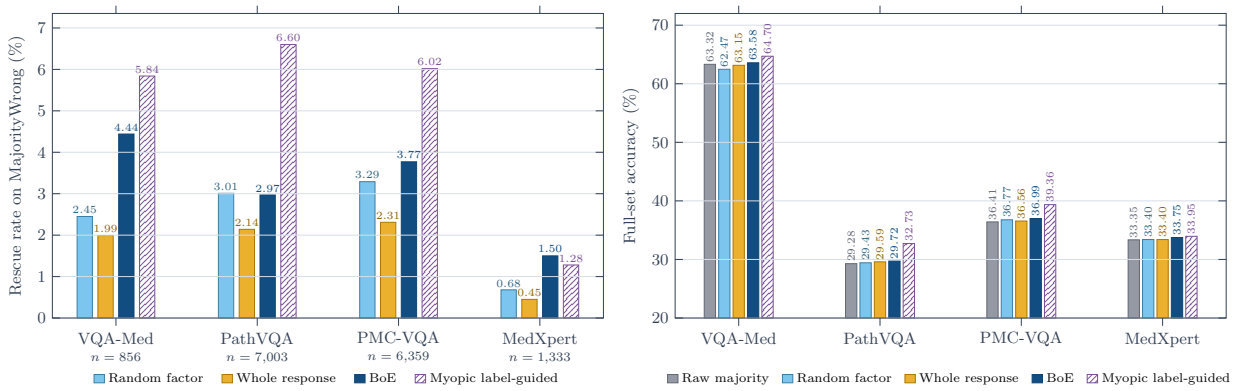

Figure 3: **Selection performance at $C = 16$. Left**, rescue rates on the policy-aligned MajorityWrong subsets; raw BoN is omitted because its accuracy is zero by definition, and subset sizes are shown below the dataset names. **Right**, full-set selected-candidate accuracy; the vertical axis begins at 20%. In both panels, the myopic label-guided allocator is a non-deployable, factor-only diagnostic with no upper-bound interpretation. BoE value labels are bolded for visual emphasis.

they reveal and how they select from the updated scores. The evidence judge returns observations but does not rank candidates.

Each open-answer candidate instantiates five slots: modality, anatomical region, view or plane, primary finding, and one answer-relevant attribute. Unresolved or inapplicable slots return $\varnothing$, and canonicalized claims are merged with stance preserved. MedXpertQA-MM additionally records the figure identifier. Structured checks have cost 1, ordinary or grounded VLM factor judgments cost 4, and whole-response judgments cost 8. These are design costs rather than measured wall-clock or token costs.

At $C = 16$, we compare raw BoN majority, random factor acquisition, whole-response judging, BoE, and a myopic label-guided allocator. The last uses evaluation labels to choose factors greedily and is a non-deployable, factor-only diagnostic rather than an upper bound. Random acquisition is also factor-only, while BoE may use whole-response checks. Their main comparison therefore uses unmatched menus; a matched replay is reported separately.

Evaluation metrics. Each dataset–channel group is assigned a discrimination index κ_g, and its categorical outcome frequencies provide the plug-in mass $\widehat{p}_e$ in Equation (11). Fitted indices, outcome frequencies, and policy evaluation use the same ledger. Open-answer correctness combines deterministic matching with semantic-equivalence grading, whereas MedXpertQA-MM uses exact matching over A–E. Because the open-answer evaluator and evidence judge use the same model tier, whole-response and channel-quality results are interpreted as mechanism diagnostics.

We report full-set selected-candidate accuracy, Oracle@K, and MajorityWrong, the questions on which raw BoN selects an incorrect candidate. Raw BoN has zero accuracy on this subset, while Oracle@K distinguishes selection-fixable cases from generation failures. Paired intervals and McNemar tests are conditional summaries of the retained ledgers and one candidate-generation seed. Besides, we do not estimate Λ_C on the VLM ledgers. For dataset d with n_d questions, policy π, budget C, and realized transcript $\mathsf{t}^{\pi}_{r,C}$ for question r, we report the entropy-shaped score proxy

$$\mathsf{Sharp}^{\pi}_{d,C} = \frac{1}{n_d} \sum_{r=1}^{n_d} \sum_{i=1}^{K} \left[H_b(s^{(0)}_{ri}) - H_b\big(\widetilde{s}_{ri}(\mathsf{t}^{\pi}_{r,C})\big) \right]. \tag{16}$$

Here $H_b(x) = -x \log_2 x - (1-x) \log_2 (1-x)$. Because $\widetilde{s}_{ri}$ is uncalibrated, H_b is only a shape function. The resulting dimensionless mean can be negative. It is neither information nor an estimate, upper bound, or lower bound for Λ_C.

5.2 Benchmark results

Figure 3 gives two views of the fixed-ledger result at $C = 16$. The right panel evaluates all questions, while the left panel restricts attention to failures of the cheap BoN decision.

Full-set selection. BoE is 0.26–0.58 percentage points above raw majority across the four cells. This consistent direction shows that evidence-aware reranking can improve a fixed candidate pool, but the magnitude is modest and does not by itself identify the allocation effect: random and whole-response policies also acquire evidence.

The closest policy comparison is therefore BoE against random factor acquisition. Table 2 shows that only the VQA-Med BoE–random interval excludes zero. PathVQA and PMC-VQA have positive BoE–raw intervals but no separated BoE–random contrast, suggesting that evidence can improve selection there without establishing an advantage for the current allocation rule.

Table 2: **Question-paired uncertainty at $C = 16$.** Intervals and gaps are in percentage points. Discordants are counts of BoE-correct/random-wrong and BoE-wrong/random-correct questions. McNemar p-values refer only to BoE versus random factor. All analyses are exploratory and uncorrected.

Dataset	BoE–raw (95% interval)	BoE–random (95% interval)	Discordants	McNemar p
VQA-Med	+0.26 [−0.47, +0.99]	**+1.11** [**+0.43**, **+1.80**]	47 / 21	0.0022
PathVQA	+0.43 [+0.05, +0.81]	+0.29 [−0.09, +0.68]	209 / 180	0.16
PMC-VQA	+0.58 [+0.19, +0.97]	+0.22 [−0.18, +0.61]	213 / 191	0.30
MedXpertQA-MM	+0.40 [−0.15, +0.95]	+0.35 [−0.20, +0.90]	20 / 13	0.30

On VQA-Med, the separated BoE–random contrast is +1.11 points, while the smaller BoE–raw interval includes zero. Because random factor acquisition is below raw majority and uses a narrower menu, this is a fixed-ledger policy contrast rather than a demonstrated absolute or pure routing gain. On MedXpertQA-MM, both paired intervals include zero, and restricting BoE to the same factor-only menu reduces the descriptive gap from +0.35 to +0.15 points.

Table 3: **Assigned channel discrimination and entropy-shaped score proxy at $C = 16$.** Each κ_g is for the dataset–channel group specified by its row and column; Structured κ_g gives the range over populated fixed-battery slots. $\mathsf{Sharp}^{\pi}_{d,C}$ is dimensionless and is not information or Λ_C.

Dataset	VLM κ_g	Grounded κ_g	Structured κ_g	Whole κ_g	$\mathsf{Sharp}^{\mathrm{BoE}}_{d,C}$	$\mathsf{Sharp}^{\mathrm{random}}_{d,C}$	$\mathsf{Sharp}^{\mathrm{whole}}_{d,C}$
VQA-Med	0.539	0.505	0.522–0.624	0.715	+1.025	+0.280	+0.653
PathVQA	0.518	0.534	0.275–0.517	0.608	−0.737	+0.092	−0.617
PMC-VQA	0.502	0.522	0.377–0.534	0.623	−0.276	−0.006	−0.202
MedXpertQA-MM	0.514	0.526	n/a	0.642	+1.006	+0.021	+0.799

Selection on raw-majority failures. The left panel of Figure 3 focuses on examples where additional evidence has room to change the BoN decision. VQA-Med gives the clearest result: BoE rescues 4.44% of raw failures, compared with 2.45% for random factor acquisition. PMC-VQA shows a smaller separation, while PathVQA is effectively tied with random acquisition.

Candidate-pool coverage limits these rescue rates. On MedXpertQA-MM, only 627 of 1,333 raw failures contain a correct candidate; the remaining 706 cannot be repaired by any selector. This supports the theoretical distinction between available evidence and generation headroom: evidence can re-rank candidates, but it cannot create a missing answer. Integer transition counts and denominator checks appear in Appendix B.

Channel discrimination and the score proxy. Table 3 relates the selection results to channel quality. VQA-Med has the strongest populated structured channel and positive BoE sharpening. PathVQA

and PMC-VQA instead have VLM indices near 0.5, some anti-correlated structured slots, and negative realized BoE proxy values. This pattern is qualitatively consistent with decision-effective reuse: evidence is useful when its channel is discriminative and its signed update can separate competitive candidates.

MedXpertQA-MM also shows why score sharpening is not sufficient for accuracy: its BoE and whole-response proxies are positive, yet their full-set changes remain small. Evidence must not only move scores, but move them in a decision-relevant direction on a pool containing a correct answer. Overall, the experiments support BoE as a fixed-pool evidence controller under favorable channels, while the remaining cells expose the weak-evidence, action-menu, and candidate-generation regimes predicted by the formulation.

6 Related Work

Candidate-level test-time selection. BoN samples several responses and selects one with a proxy[2, 3]. Process reward models add step-level supervision [5], while regularized BoN and Best-of-Tails control reward hacking and extreme-score errors [10, 32]. Other work allocates generation or verification compute adaptively and replaces pointwise scores with pairwise or criteria-decomposed judgments [4, 33, 34, 35, 36]. Their verification signal remains attached to a candidate or reasoning step. BoE treats a local claim shared by candidates with opposing stances as the verification unit.

Grounded verification and budgeted acquisition. VLM inference combines sampling with visual search and verification [37]. Tools expose local evidence through cropping, detection, retrieval, or execution [11, 20, 21]; VisualPRM, EVPV, and TIM-PRM score reasoning or verify visual premises [7, 12, 38]. ARM-Thinker, Skill-RM, OpenReward, and AgentVR organize heterogeneous evaluation resources [22, 23, 24, 25], while visual self-verification remains weak [39]. Value-of-information, active feature acquisition, and submodular methods select costly observations [40, 41, 42, 43]; group testing studies shared queries [27], factor graphs encode shared structure [26], and epistemic throughput studies complete-record verification [44]. BoE combines these threads through signed claims shared across candidates and valued by final selection.

Medical visual question answering. Medical VQA benchmarks span radiology, pathology, and biomedical figures, with large differences in answer space and difficulty. VQA-RAD introduced clinician-authored radiology questions; VQA-Med organizes modality, plane, organ, and abnormality questions; SLAKE adds semantic labels and a knowledge base [28, 45, 46]. PathVQA extends to pathology, PMC-VQA to large-scale literature figures, and MedXpertQA-MM to expert-level multimodal cases with rich clinical context [29, 30, 31]. Grounded benchmarks such as HEAL-MedVQA also measure whether answers use the relevant image region [47]. Medical answers often decompose into findings, locations, and attributes that can be checked separately. This is one domain-specific instantiation of a domain-independent partial-verification problem.

7 Conclusion

We introduced BoE, a test-time selection framework for partial verification. BoE keeps the BoN candidate pool fixed, represents reusable local claims with a signed candidate–factor graph, and allocates a limited evidence budget according to estimated selection value. Its ideal Bayesian formulation updates posterior mean utilities, while the practical implementation uses discrimination-weighted score updates and a plug-in value-of-information rule. Theoretically, we show that residual evidence capacity limits evidence-driven improvement and that shared factor queries can be substantially more query-efficient than candidate-level verification in a factor-code model. Common-ledger experiments on medical VQA support the proposed mechanism, showing that evidence is most useful when it is reliable, contrastive, and relevant to the current decision. Together, these results establish BoE as a structured extension of BoN for selection under partial verification.

Limitations and future work. Although we have proven the effectiveness of the BoE algorithm, it remains constrained by the candidate generator and the evidence it produces. If the candidate pool contains no correct answer, or if extracted claims are noisy, redundant, or incorrectly grounded, evidence acquisition cannot recover the missing information. Future work should therefore post-train generators to produce more diverse candidate pools together with canonical, contrastive, and verifiable claims. The current controller also uses a myopic plug-in EVSI-per-cost rule, which may miss complementary evidence sequences or allocate budget suboptimally. Future work should explore learned or multi-step acquisition policies, stronger calibration and stopping rules, and optimization under measured inference costs. These two directions—improving evidence-producing models and improving evidence-allocation algorithms—are complementary paths toward a stronger BoE system.

References

[1] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[2] 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.

[3] 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.

[4] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

[5] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023.

[7] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. Visualprm: An effective process reward model for multimodal reasoning, 2025.

[8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[9] Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests, 2022.

[10] Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling with minimum Bayes risk objective for language model alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9321–9347, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[11] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Junxin Wang, Dai Guan, Weijie Qiu, Zhihang Li, Yongbo Gai, Zhengyi Yang, Mengyu Zhou, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Grounding the score: Explicit visual premise verification for reliable vision-language process reward models, 2026.

[13] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics.

[14] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models, 2025.

[16] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022.

[17] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation, 2023.

[18] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.

[19] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021.

[20] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023.

[21] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022.

[22] Shengyuan Ding, Xinyu Fang, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiangyu Zhao, Haodong Duan, Xiaoyi Dong, Jianze Liang, Bin Wang, Conghui He, Dahua Lin, and Jiaqi Wang. Arm-thinker: Reinforcing multimodal generative reward models with agentic tool use and visual reasoning, 2025.

[23] Tao Chen, Gangwei Jiang, Pengyu Cheng, Siyuan Huang, Yihao Liu, Jingwei Ni, Jiaqi Guo, Mengyu Zhou, Kai Tang, Junling Liu, Qinliang Su, Xiaoxi Jiang, and Guanjun Jiang. Skill-rm: Unifying heterogeneous evaluation criteria via agent skill, 2026.

[24] Ziyou Hu, Zhengliang Shi, Minghang Zhu, Haitao Li, Teng Sun, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. Openreward: Learning to reward long-form agentic tasks via reinforcement learning, 2026.

[25] Jiazheng Zhang, Ziche Fu, Zhiheng Xi, Wenqing Jing, Mingxu Chai, Wei He, Guoqiang Zhang, Chenghao Fan, Chenxin An, Wenxiang Chen, Zhicheng Liu, Haojie Pan, Dingwei Zhu, Tao Gui, Qi Zhang, and Xuanjing Huang. Agentv-rl: Scaling reward modeling with agentic verifier, 2026.

[26] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[27] Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends® in Communications and Information Theory*, 23(1-2):1–221, May 2026.

[28] Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Conference and Labs of the Evaluation Forum*, 2019.

[29] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, Online, August 2021. Association for Computational Linguistics.

[30] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024.

[31] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding, 2025.

[32] Hsiang Hsu, Eric Lei, and Chun-Fu Chen. Best-of-tails: Bridging optimism and pessimism in inference-time alignment, 2026.

[33] Yung-Chen Tang, Pin-Yu Chen, and Andrea Cavallaro. Carbon: Calibrated best-of-n sampling improves test-time reasoning, 2025.

[34] Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Pairjudge rm: Perform best-of-n sampling with knockout tournament, 2025.

[35] 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.

[36] Shaddin Dughmi, Mahdi Haghifam, and Yusuf Hakan Kalayci. Adaptive generate-rank-verify: Inference-time search with costly verification, 2026.

[37] Ahmadreza Jeddi, Minh Ngoc Le, Amirhossein Kazerouni, Hakki Can Karaimer, Hue Nguyen, Iqbal Mohomed, Michael Brudno, Alex Levinshtein, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. Avis: Adaptive test-time scaling for vision-language models, 2026.

[38] Peng Kuang, Xiangxiang Wang, Wentao Liu, Jian Dong, and Kaidi Xu. Tim-prm: Verifying multimodal reasoning with tool-integrated prm, 2025.

[39] Mingyuan Wu, Meitang Li, Jingcheng Yang, Jize Jiang, Kaizhuo Yan, Zhaoheng Li, Hanchao Yu, Minjia Zhang, and Klara Nahrstedt. Aha moment revisited: Are vlms truly capable of self verification in inference-time scaling?, 2026.

[40] Christopher Jackson, Anne Presanis, Stefano Conti, and Daniela De Angelis. Value of information analysis in models to inform health policy. *Journal of the American Statistical Association*, 114(528):1480–1494, 2019.

[41] Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6450–6459. PMLR, 2021.

[42] Kai Wei, Rishabh K. Iyer, and Jeff A. Bilmes. Submodularity in Data Subset Selection and Active Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Proceedings*, pages 1954–1963. JMLR.org, 2015.

[43] A. Krause and D. Golovin. Submodular function maximization. tractability: Pract. 2012.

[44] Lei You. Epistemic throughput: Fundamental limits of attention-constrained inference, 2026.

[45] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[46] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021.

[47] Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A hallucination evaluation benchmark for grounded medical multimodal llms, 2026.

Appendix Contents

This appendix gives the details that are not needed for the main narrative but are useful for checking the claims. The appendix contains full proofs, the mathematical background behind the BoE, implementation details, metric definitions, additional diagnostics, and limitations.

A Formal Theory and Proofs

This appendix formalizes the access model and BoE controllers introduced in Sections 2–3, and then states and proves the theoretical results summarized in Section 4.

A.1 Formal Partial-Verification Model

The candidate-pool law, utility model, cheap context, signed incidence matrix, and candidate–factor graph are defined in Equations (1)–(5). We give the formal access and policy semantics here.

For a realized candidate pool, $\mathcal{V}_c$ indexes the K candidate nodes and $\mathcal{V}_f$ indexes the M canonical claim nodes. The edge set $\mathcal{E}_{\mathrm{cf}}$ consists of the pairs (i, j) for which $B_{ij} \neq 0$, with sign B_{ij}. Thus, the identity of a claim node is separated from candidate stance: candidates that assert and deny the same canonical proposition share one node and differ only in their edge signs.

A factor action targets a claim node and returns one acquired outcome. That outcome appears once in the transcript, although its effect may be applied to several candidates through the incident signed edges. This propagation does not create several independent observations. Candidate-level actions instead inspect a complete response and need not act through a claim node.

The graph specifies the declared verification interface. It need not explain every component of candidate utility and does not imply that $\boldsymbol{\Theta}$ fully determines $\mathbf{U}$. Candidate utility may also depend on unrepresented facts, interactions among claims, or aspects of a response that are not locally verifiable.

Let $\mathcal{A}$ be the evidence-action menu. Each realized action $e \in \mathcal{A}$ has cost $c(e) \geq c_{\min} > 0$ and can be acquired at most once; deliberate replicates must be represented as distinct menu items. Equation (6) separates view construction from the returned outcome. Fixed model or tool parameters are suppressed, and the executor seeds $\mathsf{Seed}_t^{\mathrm{exec}}$ are mutually independent and independent of the task and controller randomization.

A deterministic controller is a measurable rule on finite histories that returns either STOP or an applicable, not-yet-acquired action. A randomized controller may additionally use a private seed $\mathsf{Seed}_\pi^{\mathrm{pol}}$ independent of the task and all executor seeds. Starting from the empty history, the controller and executor induce the acquired sequence and the random number τ of acquisitions before STOP.

On the event $\{\tau \geq t\}$, let $T_t^\pi = (E_1, O_1, \ldots, E_t, O_t)$ be the random transcript prefix, with $T_0^\pi = \varnothing$. The stopped history $\overline{T}_t^\pi = T_{t\wedge\tau}^\pi$ is defined on the entire probability space and remains equal to the terminal transcript after stopping. The information available from the task and purchased outcomes after at most t acquisitions is

$$\mathcal{H}_t = \sigma(F, \overline{T}_t^\pi). \tag{17}$$

For a deterministic policy, the next stop-or-acquire decision is $\mathcal{H}_t$-measurable. For a randomized policy, it is measurable with respect to

$$\mathcal{G}_t^\pi = \mathcal{H}_t \vee \sigma(\mathsf{Seed}_\pi^{\mathrm{pol}}).$$

Its support is restricted to STOP and applicable, not-yet-acquired actions, so τ is a stopping time relative to $(\mathcal{G}_t^\pi)_t$.

Let Π_C^{acc} be the class of such access-admissible policies with $\tau < \infty$ and realized cost at most C almost surely. Their terminal action–outcome transcript is

$$T_\pi = (E_1, O_1, \ldots, E_\tau, O_\tau), \qquad c(T_\pi) = \sum_{t=1}^{\tau} c(E_t) \le C. \tag{18}$$

The controller never observes Ω, $\mathbf{U}$, the benchmark reference, or an unpurchased outcome. Equation (7) follows from the Bayes action for top-one selection. Its outer expectation averages over task instances, candidate pools, evidence outcomes, stopping behavior, and executor and policy randomization. If the supremum is attained, $\pi_C^\star$ denotes a maximizing acquisition policy.

A.2 Exact Bayesian and Ledger-Replay Controllers

Consider a policy prefix on the event $\{\tau \ge t\}$. Let $\mathtt{t} = ((e_1, o_1), \ldots, (e_t, o_t))$ be a realization of T_t^π, where $o_k \in \mathcal{O}_{e_k}$. Its realized cost is

$$c(\mathtt{t}) = \sum_{k=1}^{t} c(e_k) \le C. \tag{19}$$

We write $\mathtt{t} \oplus (e, o)$ for sequence concatenation and let $\mathcal{A}_{\mathrm{avail}}(f, \mathtt{t}) \subseteq \mathcal{A}$ contain the applicable actions that have not yet been acquired. Current-prefix statements are understood for t such that $\mathbb{P}(\tau \ge t) > 0$, under the corresponding conditional law.

Exact Bayesian controller. Under a specified observation model, Equation (8) defines the exact posterior mean. Access-admissibility implies that the policy identity and its independent randomization reveal no task information beyond F and the realized action–outcome history. The policy and prefix-length superscripts can therefore be suppressed in the notation for μ_i.

The Bayes selection is

$$i^\star(f, \mathtt{t}) \in \arg\max_{1 \le i \le K} \mu_i(f, \mathtt{t}), \tag{20}$$

with a fixed deterministic tie-breaking rule. Its current value is

$$V_{\mathrm{sel}}(f, \mathtt{t}) = \max_i \mu_i(f, \mathtt{t}). \tag{21}$$

For the empty transcript, $\mu_i(f, \varnothing) = \mu_i^{\mathrm{prior}}(f)$.

The exact EVSI in Equation (9) uses the prospective random outcome O_e and the history-conditioned predictive distribution

$$P(O_e = o \mid F = f, T_t^\pi = \mathtt{t}).$$

The corresponding one-step greedy rule is

$$e^\star \in \arg\max_{e \in \mathcal{A}_{\mathrm{avail}}(f, \mathtt{t}):\, c(\mathtt{t}) + c(e) \le C} \frac{\mathrm{EVSI}_{\mathrm{sel}}(e \mid f, \mathtt{t})}{c(e)}. \tag{22}$$

This is a sequential approximation to Equation (7), not a globally optimal acquisition policy in general. The procedure stops when no available action fits the remaining budget or when the largest value density is at most the threshold η.

Ledger-replay approximation. The ledger replay replaces the exact observation model with pooled empirical components. Each action e has a categorical outcome alphabet $\mathcal{O}_e$, and

$$\alpha_{ie} : \mathcal{O}_e \longrightarrow \{-1, 0, +1\} \tag{23}$$

records whether an outcome supports, contradicts, or supplies no applicable signal for candidate i.

For a factor action targeting claim j, this map combines the returned verdict with the graph sign B_{ij}. A confirmation uses the direction B_{ij}, a rejection reverses that direction, and an unresolved outcome or an

unrelated candidate gives zero. The same factor outcome can therefore produce positive, negative, and zero updates across the candidate pool. Candidate-level actions define $\alpha_{ie}(o)$ directly for the inspected response.

Algorithm 1 Best-of-Evidence score-based selection

Input: executor holding raw x, candidate count K, evidence budget C, action menu $\mathcal{A}$, pooled outcome masses $\widehat{p}$, discrimination indices κ, threshold η
Sample $\mathbf{y}_{1:K} \sim p_\phi^{\text{gen}}(\cdot \mid x)$ and expose the realized cheap context f
Extract canonical claims and signed candidate stances; build G_{cf}
Instantiate evidence actions with their sources, costs, and calibration groups
Initialize $\mathtt{t} \leftarrow \varnothing$ and evidence log $\mathcal{L} \leftarrow \varnothing$
while a feasible unrevealed action remains **do**
    Compute $\widehat{\text{EVSI}}_{\text{sel}}(e \mid f, \mathtt{t})/c(e)$ for every feasible action e
    Let $\widehat{e}^\star$ be the fixed-tie maximizer of the plug-in value density
    if $\widehat{\text{EVSI}}_{\text{sel}}(\widehat{e}^\star \mid f, \mathtt{t})/c(\widehat{e}^\star) \le \eta$ **then**
        break
    end if
    Execute $\widehat{e}^\star$ and observe $o^\star \in \mathcal{O}_{\widehat{e}^\star}$
    Set $\mathtt{t} \leftarrow \mathtt{t} \oplus (\widehat{e}^\star, o^\star)$
    Update all affected candidate scores using Equation (10)
    Append the acquired action and outcome to $\mathcal{L}$
end while
Return the fixed-tie maximizer of $\widetilde{s}_i(\mathtt{t})$, the score vector $\widetilde{\mathbf{s}}(\mathtt{t})$, and the evidence log $\mathcal{L}$

Let $\gamma(e)$ be the dataset–channel calibration group assigned to action e. The replay uses the pooled categorical mass

$$\widehat{p}_e(o) := \widehat{p}_{\gamma(e)}(o)$$

and the assigned selection-side discrimination index

$$\kappa(e) := \kappa_{\gamma(e)} \in (0, 1).$$

Either quantity may be estimated from the common ledger or fixed by the protocol; neither is a parameter of a fitted joint observation model. Define

$$\text{wt}(e) = \text{logit}\, \kappa(e) = \log \frac{\kappa(e)}{1 - \kappa(e)}. \tag{24}$$

Candidate scores are initialized by $\widetilde{s}_i(\varnothing) = s_i^{(0)}$, and

$$\widetilde{\mathbf{s}}(\mathtt{t}) := \big(\widetilde{s}_1(\mathtt{t}), \ldots, \widetilde{s}_K(\mathtt{t})\big).$$

All replay scores used in the reported experiments lie strictly between zero and one, so their logits are finite. Suppressing the fixed instance context f, Equation (10) applies the signed discrimination-weighted increment to every affected candidate. For a factor action, the acquired outcome and channel weight are shared, while $\alpha_{ie}(o)$ determines the candidate-specific update direction.

For each possible outcome $o \in \mathcal{O}_e$, the replay computes the hypothetical updated scores and applies Equation (11). Unlike the exact predictive distribution in Equation (9), $\widehat{p}_e$ is pooled at the dataset–channel level and is not conditioned on the current history. Consequently, $\widehat{\text{EVSI}}_{\text{sel}}$ is a plug-in action-ranking score and need not inherit exact Bayesian semantics.

The implemented greedy controller ranks all feasible actions by

$$\widehat{\text{EVSI}}_{\text{sel}}(e \mid f, \mathtt{t})/c(e),$$

selects a fixed-tie maximizer $\widehat{e}^{\star}$, and applies the same budget and threshold stopping rules as the exact controller. The hat distinguishes this implemented action from the exact-EVSI choice $e^{\star}$ in Equation (22).

Likewise, $\kappa(e)$ measures candidate discrimination rather than factor-truth accuracy or a likelihood parameter. We reserve "posterior" and "EVSI" without hats for the exact quantities in Equations (8) and (9).

The evidence log $\mathcal{L}$ records each acquired action, source, outcome, cost, assigned discrimination index, and affected candidates. A shared factor outcome is stored once. Localization metadata such as boxes, masks, and crops determines the view supplied to an action but does not enter the score update as an independent observation.

A.3 Decision-Effective Signed Reuse and Local Channel Dominance

Fix a policy π and an acquisition round t with $\mathbb{P}(\tau \geq t) > 0$. Let $h = (f, \mathtt{t})$ be a $P_{(F,T_t^{\pi})|\tau\geq t}$-almost-everywhere realization of the active history. Let $\mathbf{B} \in \{-1, 0, +1\}^{K\times M}$ be the signed candidate–claim incidence matrix and let $\mathbf{b}_j$ denote its j-th column.

Definition A.1 (Effective reuse)**.** Let $\mathbf{W}_h \succeq 0$ be a residual decision-weight matrix at history h. The effective reuse of claim j is

$$R_{\text{eff}}(j; h) = \mathbf{b}_j^{\top} \mathbf{W}_h \mathbf{b}_j. \tag{25}$$

The matrix $\mathbf{W}_h$ gives greater weight to uncertain candidates near the current decision boundary and may suppress common-mode directions. Thus, the number of candidates touched by a claim is not itself a measure of useful reuse.

Assumption A.2 (Local quadratic decision model)**.** Around history h, the decision value produced by a small candidate-score log-odds perturbation $\Delta\boldsymbol{\ell}$ has the local form

$$\Delta V \approx \frac{1}{2} \Delta\boldsymbol{\ell}^{\top} \mathbf{W}_h \Delta\boldsymbol{\ell}, \qquad \mathbf{W}_h \succeq 0. \tag{26}$$

For a factor action e targeting claim $j(e)$, let $\Delta\boldsymbol{\ell}_e = \xi_e \mathbf{b}_{j(e)}$, where $\mathbb{E}[\xi_e \mid h] = 0$, $\mathbb{E}[\xi_e^2 \mid h] = \iota_e(h)$, and the action cost is $c(e)$. For a candidate-level check on candidate i, let $\Delta\boldsymbol{\ell}_i^{\text{cand}} = \zeta_i \mathbf{e}_i$, where $\mathbb{E}[\zeta_i \mid h] = 0$, $\mathbb{E}[\zeta_i^2 \mid h] = \iota_i^{\text{cand}}(h)$, and the cost is c_i^{cand}.

Proposition A.3 (Local channel dominance)**.** *Under Assumption A.2, the expected local decision-value density of factor action e is proportional to*

Formal local comparison

$$D_{\text{factor}}(e; h) = \frac{\iota_e(h) R_{\text{eff}}(j(e); h)}{c(e)}. \tag{27}$$

The density of a candidate-level check on candidate i is proportional to

$$D_{\text{cand}}(i; h) = \frac{\iota_i^{\text{cand}}(h) \mathbf{e}_i^{\top} \mathbf{W}_h \mathbf{e}_i}{c_i^{\text{cand}}}. \tag{28}$$

Hence, within the local model, factor evidence dominates candidate-level checking when

$$\max_{e\in\mathcal{A}_{\text{fac}}(h)} D_{\text{factor}}(e; h) > \max_{1\leq i\leq K} D_{\text{cand}}(i; h). \tag{29}$$

Proof. For factor action e, substitute $\Delta\boldsymbol{\ell}_e = \xi_e \mathbf{b}_{j(e)}$ into Equation (26):

$$\mathbb{E}[\Delta V_e \mid h] \approx \frac{1}{2} \mathbb{E}[\xi_e^2 \mid h] \mathbf{b}_{j(e)}^{\top} \mathbf{W}_h \mathbf{b}_{j(e)} \tag{30}$$

$$= \frac{1}{2} \iota_e(h) R_{\text{eff}}(j(e); h). \tag{31}$$

Dividing by $c(e)$ yields $\frac{1}{2}D_{\text{factor}}(e;h)$.

Similarly,

$$\mathbb{E}[\Delta V_i^{\text{cand}} \mid h] \approx \frac{1}{2}\mathbb{E}[\zeta_i^2 \mid h]\mathbf{e}_i^\top \mathbf{W}_h \mathbf{e}_i \tag{32}$$

$$= \frac{1}{2}\iota_i^{\text{cand}}(h)\mathbf{e}_i^\top \mathbf{W}_h \mathbf{e}_i. \tag{33}$$

Dividing by c_i^{cand} gives $\frac{1}{2}D_{\text{cand}}(i;h)$. The common factor $1/2$ cancels when the two action classes are compared. □

This proposition is a local explanatory approximation to value per cost, not a universal ordering of channels and not the plug-in controller objective in Section 3.

A.4 Compressive Verification and Finite Probe Calculations

Assumption A.4 (Factor-code candidate family)**.** Let $\mathbf{\Theta} \in \{0,1\}^M$ be uniformly distributed. The candidate pool contains one candidate $\mathbf{y_a}$ for every assignment $\mathbf{a} \in \{0,1\}^M$, so $K = 2^M$. Candidate utility is

$$U_{\mathbf{a}} = \mathbf{1}\{\mathbf{a} = \mathbf{\Theta}\}. \tag{34}$$

To embed the construction in the access model, the executor holds the record $X = \mathbf{\Theta}$ and the selector starts from constant F. A factor query at coordinate j and a candidate query at assignment $\mathbf{a}$ have outcomes

$$O_{e_j^{\text{fac}}} = X_j, \qquad O_{e_{\mathbf{a}}^{\text{cand}}} = \mathbf{1}\{\mathbf{a} = X\}.$$

Both query types inspect admissible views of the same executor-held record and neither reads a utility variable or benchmark reference.

Theorem A.5 (Compressive-verification separation)**.** *Under Assumption A.4, factor-level evidence identifies the correct candidate with $M = \log_2 K$ queries and succeeds with probability one. Any policy that makes q candidate-level queries and then outputs one candidate succeeds with probability at most*

$$\min\left\{\frac{q+1}{K}, 1\right\}. \tag{35}$$

Consequently, success probability at least $1-\varepsilon$ requires at least $\lceil (1-\varepsilon)K \rceil - 1$ candidate-level queries.

Proof. Querying all M coordinates reveals $\mathbf{\Theta}$ exactly and therefore identifies $\mathbf{y}_{\mathbf{\Theta}}$. For candidate-level verification, repeated queries are never useful. Consider first $q < K$ distinct queried assignments. The probability that one of them is correct is q/K. If all queries return zero, the posterior is uniform over the remaining $K-q$ assignments, so the best final guess succeeds with conditional probability $1/(K-q)$. The total success probability is

$$\frac{q}{K} + \frac{K-q}{K}\frac{1}{K-q} = \frac{q+1}{K}.$$

For $q \geq K$, success is at most one. Solving $(q+1)/K \geq 1-\varepsilon$ gives the stated integer query lower bound. □

For direct comparison with a fixed query budget, the same candidate-query bound can be written as

$$\min\left\{\frac{q+1}{K}, 1\right\}. \tag{36}$$

The separation is a witness for possible compression through shared coordinates; it does not assert that real VLM candidate pools form complete binary codes.

Exact factor-code curves. For $K = 2^M$, after $q \le M$ exact factor queries, 2^{M-q} assignments remain possible. The optimal success probability is

$$P_{\text{factor}}(q) = \min\left\{\frac{2^q}{K}, 1\right\}. \tag{37}$$

The candidate-level curve follows from Theorem A.5:

$$P_{\text{candidate}}(q) = \min\left\{\frac{q+1}{K}, 1\right\}. \tag{38}$$

For $K = 64$ and $M = 6$, the first curve reaches one after six checks, whereas the second requires 60 checks to exceed 95% success.

Noisy-factor curve. Consider a separate model with $K = 16$, $M = 4$, and independent binary symmetric channel observations

$$O_j^{\text{BSC}} = \Theta_j \oplus N_j,$$

where $N_j \sim \text{Bernoulli}(1-\rho)$ independently across j and independently of $\mathbf{\Theta}$. Thus

$$\rho = \mathbb{P}(O_j^{\text{BSC}} = \Theta_j) \in [1/2, 1].$$

The full noisy transcript is

$$T_{\text{BSC}} = (O_1^{\text{BSC}}, \dots, O_M^{\text{BSC}}).$$

One uniform bit sent through this channel contributes $1 - H_b(\rho)$ bits, so independence gives

$$\mathrm{I}_2(\mathbf{\Theta}; T_{\text{BSC}}) = M[1 - H_b(\rho)].$$

Because the one-hot utility vector $\mathbf{U}$ uniquely identifies $\mathbf{\Theta}$ in this construction,

$$\mathrm{I}_2(\mathbf{U}; T_{\text{BSC}}) = M[1 - H_b(\rho)]. \tag{39}$$

The posterior mode is the observed code and is correct exactly when all M observed bits are correct. Relative to uniform-prior success $1/K$, its gain is

$$\Delta_{\text{mode}}(\rho) = \rho^M - \frac{1}{K}. \tag{40}$$

Both expressions vanish at $\rho = 1/2$. The parameter ρ here is a true channel correctness probability and is distinct from the empirical selection-discrimination index κ in Section 3.

A.5 Residual Evidence Capacity and Information Upper Laws

Before round t, an access-admissible policy observes only the cheap context and its purchased action–outcome history. Let Π_C^{acc} denote the policies defined in Section 2 whose realized cost is at most C almost surely.

Definition A.6 (Evidence capacity)**.** For an adaptive policy π, let

$$T_\pi = (E_1, O_1, \dots, E_\tau, O_\tau)$$

be its terminal action–outcome transcript. The residual evidence capacity is

$$\Lambda_C = \sup_{\pi \in \Pi_C^{\text{acc}}} \mathrm{I}(\mathbf{U}; T_\pi \mid F). \tag{41}$$

Conditioning on F removes information already exposed through the cheap interface. The transcript includes the adaptively selected actions as well as their outcomes, so the supremum accounts for overlap and redundancy. Because conditional mutual information is nonnegative, $\Lambda_C \ge 0$.

For the finite noisy transcript above, conversion from bits to nats gives

$$(\ln 2)\, \mathrm{I}_2(\mathbf{U}; T_{\text{BSC}}) \le \Lambda_C$$

whenever its four probes are access-admissible and fit the budget. Equality holds when those four unit-cost probes are the entire feasible menu and $C = 4$.

The access semantics are essential. If the complete raw record were included in F and each action outcome were only a deterministic function of that record plus independent noise, then the transcript would add no conditional information about $\mathbf{U}$ beyond F. The capacity definition therefore applies to the costed view-access model in Equation (6).

A.5.1 Total-information upper law

Assumption A.7 (Bernoulli candidate model)**.** Before observing cheap context or evidence, the utility vector $\mathbf{U} = (U_1, \ldots, U_K)$ has independent coordinates $U_i \sim \text{Bernoulli}(p_U)$. The cheap context satisfies

$$\mathrm{I}(\mathbf{U}; F) \le K J_F. \tag{42}$$

Equivalently, the main-text parameterization records the same assumption as

$$\mathrm{I}(\mathbf{U}; F) \le K J_F, \qquad J_F \ge 0.$$

Theorem A.8 (Evidence-capacity upper law)**.** *Under Assumption A.7, let an access-admissible, budget-feasible policy observe $\mathcal{Z} = (F, T_\pi)$ and select a subset $S(\mathcal{Z}) \subseteq \{1, \ldots, K\}$ of size m. Define*

$$T_m = \mathbb{E}\left[\sum_{i \in S(\mathcal{Z})} U_i\right]. \tag{43}$$

If J_F and Λ_C are measured in nats, then

$$T_m \le m p_U + \sqrt{\frac{m}{2}\,(K J_F + \Lambda_C)}. \tag{44}$$

If they are measured in bits, the square-root term is multiplied by $\sqrt{\ln 2}$.

Proof. Let $\mathcal{S}_m$ be the collection of all size-m subsets of $\{1, \ldots, K\}$. For fixed $s \in \mathcal{S}_m$, define

$$X_s = \sum_{i \in s} (U_i - p_U).$$

By Hoeffding's lemma,

$$\log \mathbb{E} \exp(\lambda X_s) \le \frac{m\lambda^2}{8} \qquad \text{for all } \lambda \in \mathbb{R},$$

so X_s is $m/4$-sub-Gaussian.

Let $\widehat{S} = S(\mathcal{Z})$ and, for every s with $\mathbb{P}(\widehat{S} = s) > 0$, define

$$D_s = D_{\mathrm{KL}}\left(P_{\mathbf{U}|\widehat{S}=s} \,\|\, P_{\mathbf{U}}\right).$$

The variational representation of relative entropy gives, for any $\lambda > 0$,

$$\mathbb{E}[X_s \mid \widehat{S} = s] \le \frac{D_s + \log \mathbb{E} \exp(\lambda X_s)}{\lambda} \tag{45}$$

$$\le \frac{D_s}{\lambda} + \frac{m\lambda}{8}. \tag{46}$$

Optimizing over λ yields

$$\mathbb{E}[X_s \mid \widehat{S} = s] \le \sqrt{\frac{m D_s}{2}}. \tag{47}$$

Averaging over $\widehat{S}$ and applying Jensen's inequality,

$$\mathbb{E}[X_{\widehat{S}}] \le \sum_s \mathbb{P}(\widehat{S} = s) \sqrt{\frac{m D_s}{2}} \tag{48}$$

$$\le \sqrt{\frac{m}{2} \sum_s \mathbb{P}(\widehat{S} = s) D_s} \tag{49}$$

$$= \sqrt{\frac{m}{2}\, \mathrm{I}(\mathbf{U}; \widehat{S})}. \tag{50}$$

Since $\widehat{S}$ is a function of $\mathcal{Z}$, data processing gives

$$\mathrm{I}(\mathbf{U};\widehat{S}) \le \mathrm{I}(\mathbf{U};\mathcal{Z}).$$

By the chain rule and Definition A.6,

$$\mathrm{I}(\mathbf{U};\mathcal{Z}) = \mathrm{I}(\mathbf{U};F) + \mathrm{I}(\mathbf{U};T_\pi \mid F) \tag{51}$$

$$\le KJ_F + \Lambda_C. \tag{52}$$

Finally,

$$T_m = mp_U + \mathbb{E}[X_{\widehat{S}}] \le mp_U + \sqrt{\frac{m}{2}\,(KJ_F + \Lambda_C)}.$$

The proof uses nats. If information is measured in bits, multiplying it by $\ln 2$ produces the factor $\sqrt{\ln 2}$. □

The theorem can equivalently separate the acquisition policy from its terminal selector. For any $\pi \in \Pi_C^{\mathrm{acc}}$, let δ_m^{sel} map (F, T_π) to the size-m set

$$\widehat{\mathcal{I}}_{\pi,m} = \delta_m^{\mathrm{sel}}(F, T_\pi),$$

and define

$$\mathrm{EU}_m(\pi, \delta_m^{\mathrm{sel}}) = \mathbb{E}\left[\sum_{i \in \widehat{\mathcal{I}}_{\pi,m}} U_i\right].$$

Then the main-text form is

$$\mathrm{EU}_m(\pi, \delta_m^{\mathrm{sel}}) \le mp_U + \sqrt{\frac{m}{2}\,(KJ_F + \Lambda_C)}. \tag{53}$$

This is a no-free-lunch upper bound; it does not assert that BoE attains the right-hand side.

A.5.2 Evidence-only gain

Assumption A.9 (Conditional sub-Gaussian posterior)**.** For every $F = f$ and every fixed subset s of size m,

$$X_s^{(f)} = \sum_{i \in s} (U_i - \mathbb{E}[U_i \mid F = f])$$

is $m/4$-sub-Gaussian under $P_{\mathbf{U}|F=f}$. This holds, for example, when the candidate utilities are conditionally independent Bernoulli variables given $F = f$.

Corollary A.10 (Evidence-only gain bound)**.** *Under Assumption A.9, define*

$$V_m(F) = \max_{|s|=m} \mathbb{E}\left[\sum_{i \in s} U_i \,\middle|\, F\right]$$

and

$$V_m(F, T_\pi) = \max_{|s|=m} \mathbb{E}\left[\sum_{i \in s} U_i \,\middle|\, F, T_\pi\right].$$

For every access-admissible, budget-feasible policy,

Formal evidence-only ceiling

$$\mathbb{E}\left[V_m(F, T_\pi) - V_m(F)\right] \le \sqrt{\frac{m}{2}\Lambda_C}. \tag{54}$$

The displayed form uses nats. In bits, the right-hand side is multiplied by $\sqrt{\ln 2}$.

Proof. Fix $F = f$ and let

$$S^\star(f, T_\pi) \in \arg\max_{|s|=m} \mathbb{E}\left[\sum_{i\in s} U_i \;\middle|\; F = f, T_\pi\right].$$

Write

$$\mu_s(f) = \mathbb{E}\left[\sum_{i\in s} U_i \;\middle|\; F = f\right].$$

Because $\mu_{S^\star}(f) \le \max_s \mu_s(f) = V_m(f)$,

$$\mathbb{E}\left[V_m(f, T_\pi) - V_m(f) \mid F = f\right] \tag{55}$$

$$\le \mathbb{E}\left[\sum_{i\in S^\star} U_i - \mu_{S^\star}(f) \;\middle|\; F = f\right]. \tag{56}$$

Conditionally on $F = f$, the information-selection argument from Theorem A.8 gives

$$\mathbb{E}\left[V_m(f, T_\pi) - V_m(f) \mid F = f\right] \tag{57}$$

$$\le \sqrt{\frac{m}{2}\, \mathrm{I}(\mathbf{U}; S^\star \mid F = f)} \tag{58}$$

$$\le \sqrt{\frac{m}{2}\, \mathrm{I}(\mathbf{U}; T_\pi \mid F = f)}, \tag{59}$$

where the second inequality follows by conditional data processing because $S^\star$ is a function of (f, T_π). Averaging over F and applying Jensen's inequality,

$$\mathbb{E}[V_m(F, T_\pi) - V_m(F)] \le \mathbb{E}_F \sqrt{\frac{m}{2}\, \mathrm{I}(\mathbf{U}; T_\pi \mid F = f)} \tag{60}$$

$$\le \sqrt{\frac{m}{2}\, \mathrm{I}(\mathbf{U}; T_\pi \mid F)} \tag{61}$$

$$\le \sqrt{\frac{m}{2}\Lambda_C}. \tag{62}$$

The bit-valued form follows by multiplying mutual information by $\ln 2$. □

For the terminal-history notation used in Section 4, define

$$V_m(f) = \max_{\substack{\mathcal{J}\subseteq\{1,\dots,K\}\\ |\mathcal{J}|=m}} \mathbb{E}\left[\sum_{i\in\mathcal{J}} U_i \mid F = f\right]$$

and, for P_{F,T_π}-almost every (f, t),

$$V_m^\pi(f, \mathsf{t}) = \max_{\substack{\mathcal{J}\subseteq\{1,\dots,K\}\\ |\mathcal{J}|=m}} \mathbb{E}\left[\sum_{i\in\mathcal{J}} U_i \mid F = f, T_\pi = \mathsf{t}\right].$$

Writing $V_m(F)$ and $V_m^\pi(F, T_\pi)$ for the corresponding random evaluations, the top-one notation of Section 3 satisfies

$$V_{\mathrm{sel}}(f, \mathsf{t}) = V_1^\pi(f, \mathsf{t})$$

at a terminal transcript. Equation (15) is the corresponding two-step information ceiling.

A.6 Relation to Candidate-Level JaKoB Verification

Suppose every action verifies one complete candidate, each action has unit cost, and each nonredundant verification contributes the same conditional information I_{ver}. A budget of B such actions gives

$$\Lambda_C = BI_{\text{ver}}. \tag{63}$$

With the normalization $I_{\text{ver}} = 1$, the homogeneous JaKoB form becomes

$$Bp + \sqrt{JKB} = p\Lambda_C + \sqrt{JK\Lambda_C}. \tag{64}$$

This identity relies on homogeneous, nonredundant, candidate-level checks. General BoE actions can have unequal costs, unequal discrimination, overlapping information, and shared signed effects. The product form is therefore used only as a candidate-level special-case guide, not as a general law for reusable local evidence.

B Experimental Details and Additional Diagnostics

B.1 Common Ledger and Policy Definitions

For each question, candidate generation and all potential evidence calls are performed once. The resulting offline ledger stores the candidate pool, canonical claims, signed candidate–factor graph, channel identities, action costs, evidence outcomes, and evaluation utilities. Evaluation utilities are hidden from deployable replay policies; they are used only for scoring and for the label-guided diagnostic. The evidence judge returns observations and does not rank candidates.

The main $C = 16$ comparison contains five policies. Raw BoN returns the zero-evidence majority winner. Random factor acquisition uses the same factor representation and score update as BoE but samples only factor actions. Whole-response judging purchases complete-candidate judgments. BoE ranks all available actions by the plug-in value in Equation (11). The myopic label-guided allocator uses evaluation labels at each step to reveal the factor that most improves the current selection. It is non-deployable, factor-only, and neither globally optimal nor an upper bound.

The main BoE menu includes whole-response actions, whereas random acquisition is factor-only. Their difference is therefore an unmatched-menu descriptive policy gap. A factor-only BoE replay on MedXpertQA-MM, reported below, matches the random baseline's action menu.

All four main cells use Qwen3-VL-30B-A3B for candidate generation and Qwen3-VL-235B-A22B for evidence judgment. We draw $K = 16$ candidates at temperature 1.1 and replay budgets $C \in \{1, 2, 4, 8, 16\}$. The costs 1, 4, and 8 are synthetic design costs. We did not measure wall-clock, token, or monetary costs, and the retained package does not contain a cost-ratio sensitivity analysis.

B.2 Dataset Protocols

SLAKE. The SLAKE cell contains 645 open-answer questions and uses the earlier 8B-generator/30B-judge free-factor pipeline. It does not use the fixed five-slot battery and is retained only as a historical comparison; it is not part of the four-cell main benchmark.

VQA-Med. VQA-Med provides modality, plane, organ, and abnormality questions. Because the available local copy contains only the training partition, we construct a deterministic evaluation subset of 2,334 questions. The structured modality, anatomy, and view vocabularies make this the most verifier-rich open-answer cell.

PathVQA. PathVQA contributes 9,903 open pathology questions after removing binary yes/no examples. Pathology images frequently do not support radiology-oriented view or grading slots, so those slots often return $\varnothing$ or have weak fitted discrimination. This cell measures transfer of the radiology-oriented battery to pathology images.

PMC-VQA. We sample 10,000 questions and hide the original multiple-choice options from the generator. The correct option text becomes the open-answer reference. Items whose reference is an option-specific placeholder such as "none of the above" or "cannot determine" are filtered. This protocol removes the option-term evidence channel and is not an official open-answer split.

MedXpertQA-MM. We use the 2,000-question multimodal test set. Every question is a single-answer, five-way MCQ with between one and six images. Each factor carries a figure identifier; grounded factor actions inspect the corresponding image, whereas the whole-response judge receives all images. Candidate correctness is option-count-aware exact matching over A–E. Raw BoN selects an incorrect candidate on 1,333 questions.

One MedXpertQA-MM record has no valid parsed candidate. The raw policy still selects its fallback candidate, which is evaluated as incorrect, so this item is included in the policy-aligned MajorityWrong set. We use 1,333 throughout because it exactly matches the failures of the evaluated raw policy.

Table 4: **Detailed candidate-generation statistics.**

Dataset	Questions	K	Candidate rows	Parsed (%)	Answer (%)	Oracle@K
SLAKE	645	16	10,320	90.5	98.7	0.588
VQA-Med	2,334	16	37,344	95.1	97.7	0.821
PathVQA	9,903	16	158,448	94.0	99.1	0.615
PMC-VQA	10,000	16	160,000	88.2	98.5	0.651
MedXpertQA-MM	2,000	16	32,000	96.2	94.5	0.647

B.3 Evidence Channels

The open-answer battery contains modality, anatomical region, view or plane, primary finding, and one answer-relevant attribute. The value ∅ is a first-class outcome: it marks an inapplicable or unresolved slot and contributes zero score increment. Canonicalized claims retain candidate stance. For MedXpertQA-MM, a factor also records its figure identifier.

Table 5: **Evidence channels used in the current experiments.**

Channel	Typical observation	Cost	Score role
Structured / metadata	Vocabulary, option, or metadata check	1	Channel discrimination weight
VLM factor	Whole-image factor verdict	4	Channel discrimination weight
Grounded VLM	Figure-specific factor verdict	4	Grounded discrimination weight
Whole response	Complete-candidate verdict	8	Candidate discrimination weight

Grounding metadata, masks, bounding boxes, and crop quality determine what the judge sees but do not enter the selection score as independent evidence. Grounded and whole-image VLM judgments remain separate calibration groups.

B.4 Open-Answer Evaluation and Channel Calibration

For open answers, deterministic evaluation first applies normalized string matching, substring matching, and numerical tolerance. Remaining candidate answers are graded for semantic equivalence against the reference answer. Per-example evaluation utilities are not revealed during deployable policy replay. The open-answer evaluator and evidence judge use the same model tier, so the open-answer cells do not provide an independent-evaluator test; whole-response results are consequently diagnostic. MedXpertQA-MM uses exact A–E matching and does not require an LLM correctness evaluator.

For dataset–channel group g, the selection-side discrimination index is

$$\kappa_g = 0.5 + \frac{1}{2}\left[\mathbb{P}(U = 1 \mid \text{support}, g) - \mathbb{P}(U = 1 \mid \text{contradict}, g)\right]. \tag{65}$$

This quantity is neither per-factor truth accuracy nor a likelihood-ratio parameter. It measures whether support from group g makes an affected candidate more likely to be correct. An index of 0.5 yields zero score weight, while an index below 0.5 reverses the apparent update direction. Each action inherits $\kappa(e) = \kappa_{\gamma(e)}$.

The fitted indices and categorical outcome frequencies are estimated from the same ledger used for replay. The latter define $\widehat{p}_e(o) = \widehat{p}_{\gamma(e)}(o)$ in Equation (11). The explicitly fixed Cheap values are not fitted. This protocol supports a mechanism analysis, not a held-out calibration estimate; deployment-oriented evaluation would require a separate calibration split or cross-fitting.

Table 6: **Complete assigned channel-discrimination index table.** The $s1$, $s3$, and $s5$ columns are populated fixed-battery structured channels. "n/a" means that the channel is absent. The starred Cheap value is a hand-set index in the open-answer runs.

Dataset	Cheap κ_g	VLM κ_g	Grounded κ_g	Verifier $s1$	Verifier $s3$	Verifier $s5$	Whole κ_g
SLAKE	0.55*	0.593	0.612	n/a	n/a	n/a	0.670
VQA-Med	0.55*	0.539	0.505	0.624	0.522	0.536	0.715
PathVQA	0.55*	0.518	0.534	0.517	0.310	0.275	0.608
PMC-VQA	0.55*	0.502	0.522	0.534	0.512	0.377	0.623
MedXpertQA-MM	0.505	0.514	0.526	n/a	n/a	n/a	0.642

The MedXpertQA-MM implementation carries unused hand-set defaults for battery slots that do not exist in its MCQ protocol. We report these slots as "n/a" because they do not enter its score updates.

B.5 Policy-Specific Score-Proxy Diagnostics

Equation (16) is evaluated separately for each replay policy. Its columns are not additive channel components, and negative values are permitted because the underlying candidate scores are uncalibrated.

Table 7: **Policy-specific entropy-shaped score proxy at $C = 16$.** Values are dimensionless means of $\mathsf{Sharp}^{\pi}_{d,C}$ under separate replay policies. The label-guided factor column is a non-deployable diagnostic and has no upper-bound interpretation.

Dataset	BoE policy	Whole response	Grounded factor	Label-guided factor	Random factor
VQA-Med	+1.025	+0.653	+0.011	+0.193	+0.280
SLAKE	+0.607	+0.389	−0.003	+0.142	+0.235
MedXpertQA-MM	+1.006	+0.799	+0.021	+0.015	+0.021
PMC-VQA	−0.276	−0.202	−0.020	+0.027	−0.006
PathVQA	−0.737	−0.617	−0.116	+0.297	+0.092

On PathVQA, the label-guided factor diagnostic has a positive proxy value while the realized BoE update does not. On MedXpertQA-MM, whole-response replay has a large positive proxy value although its accuracy is nearly unchanged from raw majority. These cases show why $\mathsf{Sharp}^{\pi}_{d,C}$ should not be read as information or accuracy gain.

B.6 Paired Inference and Denominator Integrity

The main text reports question-level paired-bootstrap 95% intervals and exact two-sided McNemar tests for BoE versus random factor acquisition. The retained summaries do not preserve the bootstrap resample count, interval construction, or random seed. They are therefore conditional fixed-ledger

summaries rather than independently reproducible estimates. The analyses are exploratory, uncorrected for multiple comparisons, condition on one candidate-generation seed, and omit candidate-pool and calibration-estimation variability.

The full-set accuracy change is reconstructed from question-level transitions:

$$\Delta\text{Acc} = \frac{N(\text{raw wrong, BoE correct}) - N(\text{raw correct, BoE wrong})}{N}. \tag{66}$$

This identity keeps corrections and harmful flips on a common denominator; subtracting separately normalized conditional rates is not an accuracy change.

Table 8: **Integer reconstruction of BoE–raw accuracy at $C = 16$.** Corrections are raw-wrong/BoE-correct transitions; harms are raw-correct/BoE-wrong transitions.

Dataset	N	Raw correct	Raw wrong	Corrections	Harms	ΔAcc (pp)
VQA-Med	2,334	1,478	856	38	32	+0.26
PathVQA	9,903	2,900	7,003	208	165	+0.43
PMC-VQA	10,000	3,641	6,359	240	182	+0.58
MedXpertQA-MM	2,000	667	1,333	20	12	+0.40

For MedXpertQA-MM, the 2,000 questions decompose into 667 raw-correct questions, 627 raw-wrong questions whose pool contains a correct candidate, and 706 raw-wrong questions whose pool does not. BoE repairs 20 of the 627 fixable failures and harms 12 of the 667 raw-correct questions, yielding $(20 - 12)/2000 = +0.40$ percentage points. Random factor, whole-response, and the myopic label-guided allocator repair 9, 6, and 17 of the 1,333 raw failures, respectively.

B.7 Matched-Menu and Budget Diagnostics on MedXpertQA-MM

The main BoE policy can purchase whole-response judgments, whereas random acquisition is factor-only. A separate replay removes whole-response actions from BoE so that the action menus match.

Table 9: **MedXpertQA-MM action-menu comparison at $C = 16$.** The menu-matched replay comparison is BoE factor-only versus random factor.

Policy	Accuracy (%)	Gap from random factor (pp)
Raw BoN	33.35	—
Random factor	33.40	0.00
BoE factor-only	33.55	+0.15
BoE full menu	33.75	+0.35

The available matched-menu gap is only three questions and has no reported paired interval. It should therefore be treated as descriptive. The full-menu gap cannot be interpreted as a pure routing effect.

The existing ledger also supports a budget sweep without regenerating candidates. Table 10 shows no monotone increase in the BoE–random gap. Corrections and harms refer to BoE versus raw BoN.

Table 10: **MedXpertQA-MM fixed-ledger budget sweep.** Accuracies and gaps are percentages and percentage points, respectively.

C	Raw BoN	BoE	Random	BoE–random	Corrections	Harms	BoE–raw
1	33.35	33.35	33.30	+0.05	5	5	+0.00
2	33.35	33.30	33.20	+0.10	6	7	−0.05
4	33.35	33.55	33.20	+0.35	9	5	+0.20
8	33.35	33.45	33.05	+0.40	15	13	+0.10
16	33.35	33.75	33.40	+0.35	20	12	+0.40

Across $C \in \{1, 2, 4, 8, 16\}$, BoE–random remains at or below 0.40 points and does not grow monotonically. Together with the $C = 16$ paired interval in Table 2, this does not support a stable MedXpertQA-MM allocation gain.