CONFIDENCE IS NOT ENOUGH: DECISION-VALUE FINDING SELECTION BEATS CONFIDENCE-GATED TEST-TIME COMPUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time compute methods decide how much extra computation to spend by reading the model's own confidence. We show that signal is invalid exactly when the model lacks the information it needs to answer: on interactive MedQA-USMLE the confidence–correctness correlation is near zero under minimal information and becomes predictive only as the record is read. $16\times$ self-consistency at low information buys nothing ($+0.017$, $p{=}0.72$) while revealing two findings gains $+0.197$ ($p{<}2{\times}10^{-4}$), even beating a gold-aware compute allocator ($+0.092$, $p{=}0.044$). We make finding selection first-class and control it by the classical *decision* value of information: the finding most likely to flip the model's greedy answer. The criterion is training-free, gold-free, and survives miscalibration because a flip is a behavioral event. The prevailing selection criterion, greedy entropy reduction (the literature's "EIG"), is empirically indistinguishable from confidence gain and inherits the same pathology. Decision-value and entropy-reduction coincide under calibration and diverge under insufficiency, at two layers—selection and stopping—with which layer pays off set by *how* the belief is miscalibrated: selection wins under rank miscalibration ($+0.13$–0.16), stopping under level miscalibration, where the continuous signal saturates (72–95% of states $< 10^{-3}$) but the binary flip transfers ($+0.05$–$+0.13$ across Qwen2.5 scales, $+0.24$/$+0.29$ on reasoning models, at matched reading budget). One decision-value currency arbitrating read/compute/stop Pareto-dominates the deployed controller ($+0.255$, $p{<}2{\times}10^{-4}$) and clears the entire pure-reading frontier, beating its best point by $+0.041$ at 45% less budget. *Scope:* four-way multiple-choice reading over an available record, against a self-consistency compute baseline; we do not evaluate acquisition of genuinely unobserved outcomes, nor compute controllers built on verifiers, process rewards, or search.

1 INTRODUCTION

A growing family of methods improves LLM accuracy by spending more computation at inference—self-consistency, budget forcing, uncertainty-gated adaptive spending (Wang et al., 2023; Muennighoff et al., 2025; De Sabbata et al., 2024)—and the frontier is now adaptive allocation itself: compute-optimal routing by estimated difficulty (Snell et al., 2024), constrained-policy budget allocation (Zhai et al., 2026), uncertainty-gated verifier calls (Qu, 2026), and *learned* self-assessment read from the forward pass (Manvi et al., 2025; Chen et al., 2026). The signals differ and we do not claim they are all confidence in disguise; what they share is the argument of the function. Every one is computed from the model's own state given a *fixed* input and decides only *how much more to spend on that input*—the property we attack, and it does not care how sophisticated the read-out is.

Figure 1 summarises the argument. That signal is miscalibrated exactly where it matters most: *when the model does not yet have the information required to answer.* A clinician-facing agent diagnosing from a one-line complaint, or a QA agent given a question before the passage, is not in a state where "think harder" helps—the missing facts must be *revealed*. Under insufficiency a model's confidence is nearly uncorrelated with its correctness, so a confidence-gated controller cannot distinguish "reason more" from "read more" and predictably wastes compute (§3).

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.

The fix is to make information selection first-class—let the agent choose *which* piece to reveal next—and the question is by what criterion. The LLM information-gathering literature has converged on greedy entropy reduction, the "expected information gain" (EIG) rule of Bayesian experimental design. We show this is a confidence-family criterion (§4) and inherits the same pathology: where the model is confidently wrong, entropy-reducing selection merely *sharpens the wrong belief*. We select on the *decision* instead—information has value only insofar as it changes it (Howard, 1966)—with VOI-Δ, which scores a candidate by whether it flips the greedy answer. It is a *switch detector*, not an estimator of Howard's expected-utility quantity, a distinction made precise in §4.3.

Contributions.

1. **A mechanism, and what it rules out.** Confidence validity collapses under information insufficiency and recovers as the record is read, across three scales and two confidence estimators; a head-to-head on the same testbed shows that adding compute does not substitute for the missing facts (§3).
2. **A criterion, and the two layers it splits into.** We make the decision the object of selection, show this is not equivalent to entropy reduction, and give a training-free, gold-free switch detector, VOI-Δ (§4). Which layer pays off is set by *how* the belief is miscalibrated (§4.2): under rank miscalibration VOI-Δ is at or above fixed order, random order and ER from two findings on, across 7B/14B/32B, RACE and Llama-3.1-8B (§5.1); under level miscalibration stopping gains, carried by the continuous marginal decision value on a calibrated belief and by the ordinal flip on LLM beliefs, beating confidence-gated stopping at matched budget (§5.2). No monotone recalibration closes that gap (Prop. 3).
3. **One currency, and a tested boundary.** A single decision-value rule arbitrates read/compute/stop and Pareto-dominates the deployed confidence-gated controller ($+0.255$, $p<2\times10^{-4}$) with no stopping threshold and no reading-budget parameter, clearing the pure-reading frontier by routing a minority of steps to compute; realizable selectors and learned stoppers all halt below the gold-aware oracle (§5.4).

Scope. Decision focus improves *selection* where the belief's *ordering* is miscalibrated (the interactive-LLM regime, an ordering problem over an available record), and *stopping* where the belief's *level* is miscalibrated—the continuous magnitude needs a level-calibrated belief (DDxPlus supplies this, an LLM does not), while the ordinal flip carries stopping to LLMs. We claim a mechanism demonstrated in a controlled environment for the magnitude and on two interactive-LLM benchmarks for the flip, with a predictive boundary (ρ) for transfer—not a universal free lunch.

2 Related Work

Test-time compute and adaptive allocation. Self-consistency (Wang et al., 2023), budget forcing (Muennighoff et al., 2025), and rational metareasoning (De Sabbata et al., 2024) allocate computation from an uncertainty signal. A substantial line makes that axis adaptive—per-prompt allocation, constrained-policy budgeting, verifier-call gating, process-reward pruning, bandit difficulty estimation, granularity regulation (Snell et al., 2024; Zhai et al., 2026; Qu, 2026; Bilal et al., 2026; Zuo & Zhu, 2025; Krasniqi et al., 2026)—with scaling laws (Chen et al., 2025; Zhang et al., 2025). These are stronger baselines than fixed self-consistency and we do not claim to dominate them on their own turf. Two facts separate them from our setting: their control signals are all functions of the output distribution over a *fixed* input, so §3 applies unchanged; and they evaluate on MATH/GSM8K/AIME, where the needed information is already in the prompt—the regime our claim is confined to is the one those benchmarks exclude by construction.

Controllers that arbitrate read vs. compute, and decision-theoretic VOI. Closest to our §5.4 is RTTC (Muñoz & Yuan, 2025), which selects per query between retrieval and lightweight test-time training via a pretrained reward model. It shares our premise that reading and computing compete for one budget, but needs that reward model and has no confidence-validity analysis or stopping layer. Howard (Howard, 1966) established that information has value only when it changes the optimal decision; we borrow that as *motivation*, not equivalence: classical decision-VOI is an expected Bayes-risk improvement requiring a utility model, whereas our estimator detects a realized answer switch and drops that factor (§4.3). Gans (Gans, 2026) frames this gap; VOI-Δ carries no utility model under 0–1 loss.

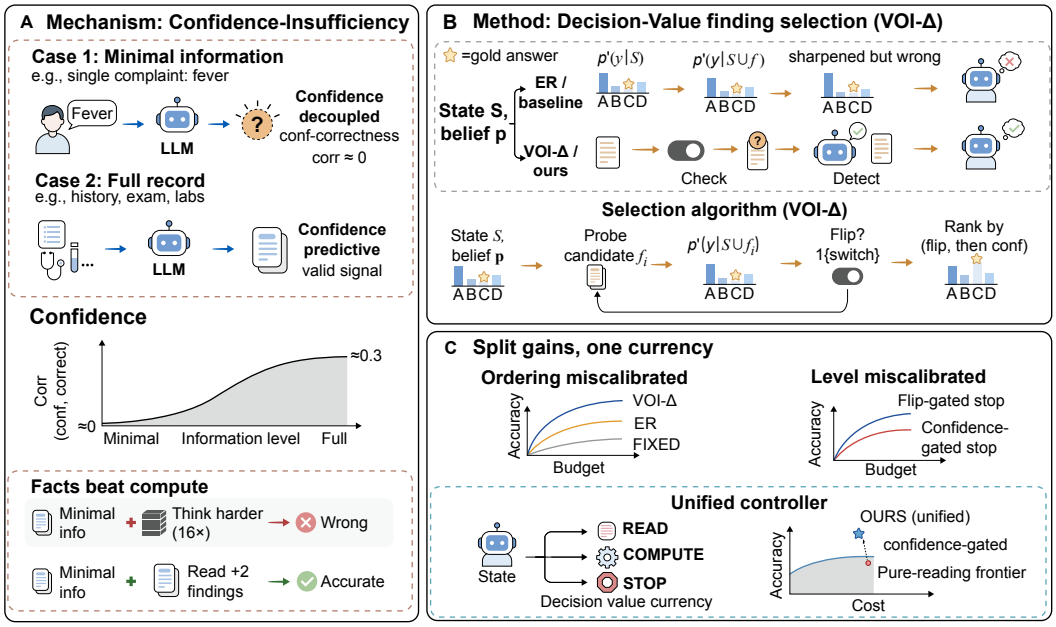


Figure 1: **The mechanism, the criterion, and one currency.** (**A**) Under minimal information confidence is decoupled from correctness and becomes predictive only once the record is read, so extra compute cannot substitute for the missing facts. (**B**) Every one-step-lookahead criterion probes the same candidates: entropy reduction sharpens the belief, a wrong one included, while VOI-Δ detects whether the greedy answer switched, ranking candidates by the flip and breaking ties on post-reveal confidence (★ = gold answer). (**C**) Decision focus pays off at selection under ordering miscalibration and at stopping under level miscalibration; one currency then arbitrates read, compute and stop.

LLM information gathering. BED-LLM (Choudhury et al., 2025) selects queries by greedy entropy reduction, in the Bayesian experimental-design tradition (Lindley, 1956; Chaloner & Verdinelli, 1995). Our hardest comparison: implemented in the testbed, it selects the *same finding as confidence* at every state we observed (Claim 1), so it is a confidence-family criterion, and decision-change selection reaches first-correct with ∼30% fewer findings while dominating it on whole-curve AUC. Amortising the objective (Hartmann et al., 2026) sharpens this: a policy trained on the entropy-reduction reward inherits the pathology. Recent VOI/clarification-agent work (Dong et al., 2026; Suri et al., 2025; Peng et al., 2026) instantiates value as information or utility and assumes a reward model; none analyze the decoupling under insufficiency or show entropy reduction ≡ confidence gain.

Calibration, selective prediction, and active feature acquisition. Neural confidence is miscalibrated (Guo et al., 2017), and a focused line shows the same for LLMs on medical questions (Savage et al., 2025; Bentegeac et al., 2025; Wu et al., 2024; Martinez, 2026; Xiong et al., 2024; Tian et al., 2023; Groot & Valdenegro-Toro, 2024). Selective prediction abstains under low confidence (El-Yaniv & Wiener, 2010; Mozannar & Sontag, 2020). We differ in *kind*: it asks whether to *abstain* given a fixed input; we ask whether the low-confidence state is *information-starved* or *compute-starved*, and show confidence cannot distinguish these while decision value can. Active feature acquisition (Aronsson et al., 2025; Valancius et al., 2023; Rahbar et al., 2023) gives the unified POMDP statement (acquire-vs-stop) that our decomposition reaches from the classical side; we add *why* that belief is untrustworthy for LLMs (§3) and the resulting split of where decision-focus pays off.

3 The Confidence-Insufficiency Mechanism

Testbed. We turn MedQA-USMLE into an interactive reading problem over an available record: the vignette is split into per-sentence *findings*; the agent observes the question, options, and a subset

of findings. "Minimal information" reveals none; "full information" reveals all. Confidence is estimated two ways, to rule out a proxy artifact; all numbers use live vLLM with real answer keys, and the per-estimator decoding is in the table note.

Result. Table 3 gives $\mathrm{corr}(\mathrm{confidence}, \mathrm{correctness})$ by information level.

Under insufficiency a confident model is barely more likely to be correct than an unconfident one (7B: $\mathrm{acc}|\mathrm{conf}{\geq}0.8 = 0.411$ vs 0.408); with full information confidence sharply separates correct from wrong. The two sides of the correlation therefore use different decoding, but that cannot manufacture the gap: the identical procedure yields $+0.293$ at full information, and a mismatch severe enough to null the minimal-information correlation would null that one too. The first-token log-probability estimator, whose two sides come from one forward pass, gives the same pattern. *Honest nuance:* minimal-info validity *rises* with scale ($0.023{\to}0.085{\to}0.147$) but stays far below full-info validity at every scale.

Compute cannot substitute for the missing facts (the title claim; Fig. 1A). If confidence is invalid under insufficiency, so is the compute axis it gates. We test that directly on the same testbed, crossing information level with compute budget $m \in \{1, 4, 8, 16\}$ (self-consistency majority vote over m temperature-0.7 samples; full grid in Table 8).

Compute at minimal information buys nothing ($+0.017$ $[-0.060, +0.094]$, p=0.72); at full information it buys $+0.077$ $[+0.009, +0.145]$ (p=0.040). Revealing two findings with a single pass beats $16\times$ compute-at-minimal by $+0.179$ $[+0.085, +0.274]$ (p=0.0004) and beats a grid-restricted gold-aware allocator of that budget by $+0.092$ (p=0.044; Appendix C); on the larger n=300 draw of §3, adding compute at minimal information moves accuracy $0.410{\to}0.410$. The per-case structure behind that bound is in Appendix C. *Scope:* the claim is about *this* compute procedure on *this* testbed—we do not test verifier-guided search, process-reward allocation, or retrieval—and the compute×info *interaction* does not resolve at this sample size (MDE ≈ 0.09), so the conclusion rests on the endpoint contrasts, phrased as *substitutability*, not a signed interaction.

4 Method: Decision-Value Finding Selection

4.1 Principle: Decision-VOI, not Information-VOI

The two criteria are contrasted in Fig. 1B. Let the agent hold belief $p := p(y \mid S)$ over answers $y \in \{A, B, C, D\}$ given revealed set S, with greedy decision $\hat{y}(p) = \arg\max_y p_y$ and gold y^*. Revealing finding f maps $p \to p' := p(y \mid S \cup f)$.

- **Entropy-reduction selection (ER).** Pick $\arg\max_f [H(p) - H(p'_f)] = \arg\min_f H(p'_f)$—the finding that makes the model most certain. This is the "EIG" rule of the information-gathering literature; on our interactive-LLM testbed the candidate's content is rendered into the state *before* the score is read, so the quantity realized is a *reduction of entropy after observing the content*, not the *expected* information gain a Bayesian experimental design computes over an observation model. We keep the honest name "entropy reduction" (ER) for the LLM criterion and reserve "EIG" for the DDxPlus environment, where a Bayesian answer model makes it a genuine expectation (§5.2, Appendix E).
- **Decision-VOI (Howard, 1966):** with 0–1 loss, $VOI(f) = \mathbb{E}[\,\mathbf{1}\{\hat{y}(p'_f) \neq \hat{y}(p)\} \cdot$ value-of-switch $]$—the finding that changes the chosen answer, weighted by how much the change is worth. Our estimator addresses only the first factor (§4.3).

Claim 1 (entropy-reduction $\approx$ confidence gain—an empirical near-equivalence). Selecting to minimize posterior entropy and selecting to maximize top-class probability induce nearly the same ordering. On the 4-simplex this is not an identity, so we state it empirically and measure it at the level that matters for selection—the per-state ordering over candidate findings: on all six Qwen2.5 interactive-LLM runs, ER and VOI-conf select the *identical* finding at every observed state (selection-agreement Wilson lower bound ≥ 0.968), and on Llama-3.1-8B agreement is 88.0% $[80.9, 92.7]$—the one divergence the geometry predicts. As corroboration, across all $\approx 1.5{\times}10^5$ DDxPlus belief states the pooled rank correlation between $-H$ and $\max_y p_y$ is Spearman ρ=0.9922. The near-equivalence has an elementary geometric explanation: writing $m = \max_y p_y$ and q the residual over the other $K{-}1$ classes, $H(p) = H_b(m) + (1 - m)H(q)$; the

first term orders candidates exactly as confidence, and the second is bounded by $(1-m)\log(K-1)$, so disagreement closes linearly in $1-m$ (full decomposition in Appendix A).

Observation 2 (the flip is necessary for correction, and that is all it is). An elementary identity, stated as an observation: it does *not* show that maximizing flip probability maximizes correction probability, because $P(\text{correct} \mid \text{flip}, f)$ may vary across findings. Fix a wrong state, $\hat{y}(p) \neq y^*$. If $\hat{y}(p'_f) = \hat{y}(p)$ then $\hat{y}(p'_f) \neq y^*$, so $P(\text{correct} \mid \text{no flip}) = 0$ exactly; hence $P(\text{correct}) = P(\text{flip}) \cdot P(\text{correct} \mid \text{flip})$. Entropy-reduction selection targets sharpening (a confidence-family objective, satisfiable by sharpening a wrong belief without moving it); decision-VOI targets the flip directly and maximizes the *first* factor. *What this does not say:* VOI-Δ scores every switching candidate 1 and ties by post-confidence, so it may select a flip to a *different wrong* option; the second factor is measured to be essentially criterion-independent (0.68 vs 0.70, §5.1), so $\sim$95% of the gain comes from raising $P(\text{flip})$.

4.2 Two Layers, Two Miscalibration Regimes

Where each layer gains. Observation 2 governs *selection*, whose gain is largest when the belief's *ordering* is miscalibrated (the LLM regime); at a calibrated argmax, flips are rare and mostly harmful, and selection has no room to help (§5.2). Decision-focus re-enters at *stopping*: a rule compares the marginal decision value of continuing against a threshold. Confidence-stopping halts when $\text{conf}(S) \geq \tau$; even a rank-calibrated belief can be *level*-miscalibrated, so confidence-stopping halts precisely where confidence is most inflated. Decision-value stopping halts when $\mathbb{E}_a[\max_y p(y \mid S \cup \{f{=}a\})] - \max_y p(y \mid S)$ falls below ε—a *marginal* quantity robust to a constant level bias.

Proposition 3 (what recalibration can and cannot change). (i) *Selection is invariant under elementwise recalibration*: any strictly increasing g applied elementwise and renormalized (temperature scaling for $g(p){=}p^{1/T}$) preserves coordinate ordering, hence the argmax and the flip signal. (ii) *The confidence-stopping frontier is invariant under monotone transforms of the stopping score*: replacing $\text{conf}(S)$ by $h(\text{conf}(S))$ for strictly increasing h leaves the accuracy-vs-budget Pareto curve pointwise unchanged, so no re-thresholding (Platt (Platt, 1999), isotonic, multiclass (Zadrozny & Elkan, 2002)) can improve confidence-stopping. (iii) *Multiclass recalibration is not such a transform*: $\widetilde{\text{conf}}(S) = g(p_{\max})/\sum_j g(p_j)$ reorders states and does move the frontier—measured in §5.2, the *wrong* way (-0.03 to -0.04). **Corollary.** Stopping can only be improved by a signal outside the monotone-transform family, and it does so iff $\rho(\Delta\text{dV}, \Delta\text{acc})$ is high—a measurable ex ante predictor of transfer ($\rho{\approx}0.47$ on DDxPlus, transfers; magnitude $\rho{\approx}0.09$ on the LLM, does not; flip $\rho{=}0.24$–0.64, does). The invariance needs h *strictly* increasing and a function of $\text{conf}(S)$ alone (Appendix B).

Proposition 4 (temperature invariance: the flip is ordinal, confidence and entropy reduction are cardinal). The binary flip $\mathbf{1}\{\arg\max p \text{ changes}\}$ is invariant under any monotone recalibration, in particular $p_T \propto \exp(\text{logit}/T)$ for every $T{>}0$. Confidence and entropy reduction are *scale* quantities that degenerate as the belief sharpens ($\max_y p_T \to 1$ as $T \to 0$, $\to 1/K$ as $T \to \infty$). A reasoning model trained to commit sits at the $T{\to}0$ end: its confidence is ≈ 1.0 on every state and carries no rank information, while the flip carries the same information as anywhere else (Fig. 8, appendix).

4.3 The Algorithm (VOI-Δ)

VOI-Δ (Fig. 1B, lower) ranks candidates by the lexicographic pair $(\,\mathbf{1}\{\hat{y}(S{\cup}f) \neq \hat{y}(S)\},\ \text{conf}(S{\cup}f)\,)$ and reveals the maximizer under a budget k: the primary key is the flip indicator; the secondary key is the *post-reveal* confidence. When no candidate flips, the rule degenerates exactly to VOI-conf—a defined fallback, not a hidden confidence dependence. VOI-Δ is a *switch detector*, not a plug-in estimator of decision-VOI: it drops the value-of-switch factor, evaluates the switch at a single realized outcome, and is myopic; in the wrong-answer regime a switch is necessary for correction (Obs. 2), so maximizing switch probability maximizes the only factor a gold-free signal can see. **VOI-hybrid** uses confidence for the cold-start first pick and decision-VOI thereafter. **Cost.** Scoring one step ahead costs one probe per remaining candidate per step; *every* one-step-lookahead criterion (ER, VOI-conf, VOI-Δ) pays the identical probe bill, so our VOI-Δ-vs-ER comparison

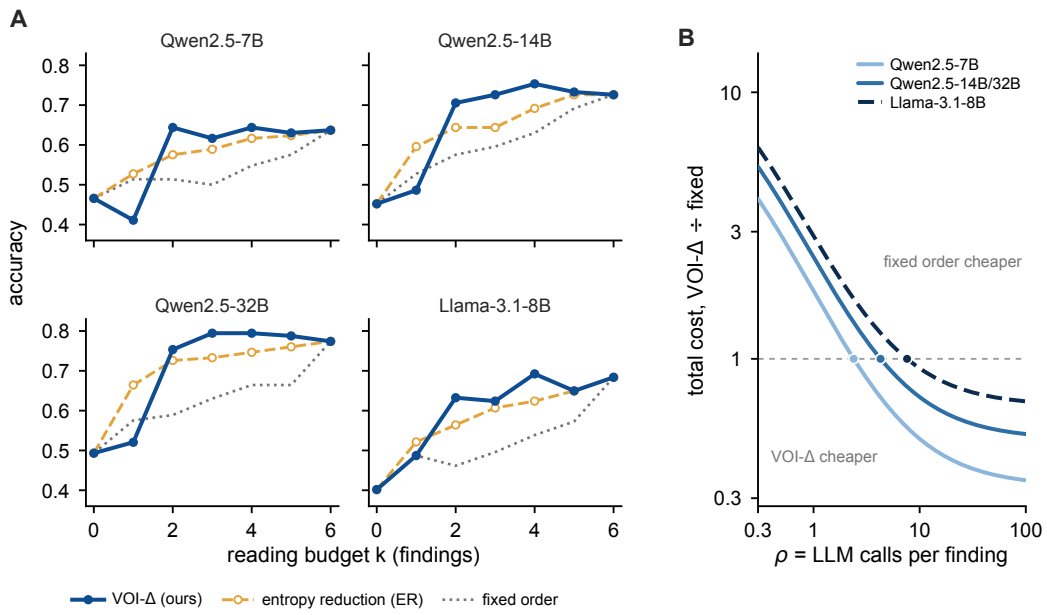


Figure 2: **Decision-value selection, and the exchange rate it needs.** **A**, Accuracy against the reading budget k on the MedQA testbed, one panel per model on a shared scale; n=146 for Qwen2.5, n=117 for Llama-3.1-8B. From k=2 on, VOI-Δ (navy, solid) is at or above entropy reduction (amber, dashed) and fixed order (grey, dotted) in every panel; at k=1 its cold start falls below both on Qwen2.5, which VOI-hybrid removes. Point estimates; per-budget tests in Appendix C. **B**, With one LLM call as the unit and ρ the cost of revealing one finding, the curve is the total cost VOI-Δ needs to reach fixed order's full-record accuracy, divided by what fixed order needs; below one, decision-value selection is the cheaper way to buy that accuracy. The break-even is $\rho \approx 2.4$–7.6 (dots). 14B and 32B reach the target at the same k^*=3 with the same probe count, so their cost curves are identical and are drawn as one.

is probe-matched. Only FIXED and RANDOM are probe-free; against them we report both axes (Appendix C.1). VOI-Δ's probes need only a single greedy token, so under a matched call count its calls are strictly the weaker requirement.

5 Experiments

Setup. MedQA-USMLE (Jin et al., 2020) and RACE (Lai et al., 2017) as interactive reading problems; DDxPlus (Tchango et al., 2022) as a controlled calibrated environment; MediQ (Li et al., 2024) as a negative headroom control (§5.2). Models are Qwen2.5 (Yang et al., 2024) and Llama-3.1 (Grattafiori et al., 2024), served with vLLM (Kwon et al., 2023). Baselines: FIXED (vignette order), RANDOM, ER (BED-LLM criterion), VOI-conf (the mechanism control), ORACLE (gold-aware). Paired bootstrap 5000 throughout; full protocol in Appendix E.

5.1 Selection: VOI-Δ beats order baselines and ER from two findings on

MedQA testbed, findings capped at 6 (n=146 for Qwen2.5, n=117 for Llama). Figure 2A gives the whole accuracy-versus-reading-budget curve rather than a single budget, because the comparison is a statement about the curve.

ER and VOI-conf selected the *same finding at every observed state* on the three Qwen2.5 runs, and agreed on 88.0% of states on Llama-3.1-8B (Claim 1), so the SOTA finding-selection criterion behaves as a confidence-family signal. VOI-Δ beats ER on accuracy and, most robustly, on efficiency—*findings-read-to-first-correct* (lower better): 7B 1.43 vs ER 1.97; 14B 1.31 vs 1.79; 32B 1.03 vs 1.50 ($+0.53$/$+0.49$/$+0.47$ fewer findings, all $p<2\times10^{-4}$). Whole-curve AUC: VOI-hybrid wins every scale. *Less is more*: reading the right findings and stopping matches reading everything

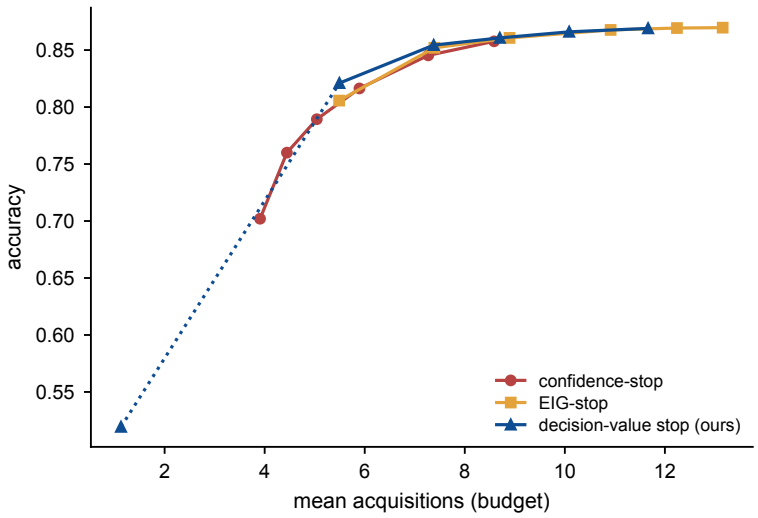


Figure 3: **Stopping on the calibrated DDxPlus belief.** Accuracy against acquisition budget as each stopping threshold is swept on one shared EIG trajectory (n=3000 patients, paired; matched-budget tests, n=10^4, in Table 4). Decision-value stopping leads in the mid-budget regime, where confidence stops early at inflated confidence; the advantage narrows with budget and reverses at aggressive budgets, which is why we report a regime rather than dominance. Dotted: no decision-value operating point between 1.1 and 5.5 acquisitions.

(32B VOI-Δ at k=3 reads 2.95 findings for 0.795 against 0.774 from all 5.28; the +0.021 difference does not separate from zero, p=0.26).

The divergence, measured directly. At every sampled wrong-answer state (MedQA-7B, 700 paired states) we compare which finding each criterion picks. ER corrects only 20.6% of the time (it sharpens a wrong belief); VOI-Δ corrects 37.3%, +0.167 at identical reading cost ($p<2\times10^{-4}$; replicated at 14B: +0.153). The conditional rate $P(\text{correct} \mid \text{flip})$ is essentially criterion-independent (0.68 vs 0.70), so +0.159 of the +0.167 (∼95%) comes from raising $P(\text{flip})$, and only +0.008 from choosing better among flipping candidates (full table in Appendix C).

Generality: a non-clinical domain and a second family. RACE (7B/14B/32B) and Llama-3.1-8B replicate the full chain—mechanism correlation, ER/VOI-conf agreement, and VOI-Δ efficiency (RACE +0.23–0.28 over ER, Llama divergence +0.161)—with the *accuracy* margin shrinking where minimal-info confidence is more valid, exactly as the mechanism predicts (Appendix C).

5.2 Stopping Under Level Miscalibration

A controlled calibrated endpoint (DDxPlus). A partial-observation classifier over 49 pathologies supplies the belief; a Bayesian answer model makes EIG a genuine expected-information-gain competitor. Acquisition raises accuracy 0.06→0.997; all rules run on the *same* EIG trajectory, differing only in the stopping signal (paired, n=10^4). The belief is order-calibrated (ECE 0.034) but *level-drifts* (mean confidence−accuracy $-0.073\rightarrow+0.088$, changing sign mid-trajectory). **Selection** does not help here (EIG 0.870 > decision-value 0.710, the calibrated endpoint Obs. 2 predicts), but **stopping** does: decision-value stopping beats confidence- and EIG-stopping in the mid-budget regime (+0.014, $p<2\times10^{-4}$; the advantage narrows and reverses at aggressive budgets, Fig. 3), and survives temperature scaling that halves validation ECE—the one reordering channel Prop. 3(iii) admits moves the frontier the *wrong* way (−0.03 to −0.04; Appendix B).

When does decision-value stopping transfer? The flip transfers, the magnitude does not. Our first two attempts returned nulls—MediQ (no reading headroom) and MedQA (headroom, but the magnitude collapsed to $\rho\approx0.09$). The continuous marginal decision value saturates on peaked first-token beliefs ($< 10^{-3}$ in 72–95% of states); the binary flip $\mathbf{1}\{\arg\max \text{ changes}\}$ does not (flip > 0 in 20–33%) and carries rank information (ρ=0.24–0.64 vs 0.086). Flip-gated reading on two interactive-LLM domains with genuine headroom:

Table 1: Flip- versus confidence-gated stopping. *Self-term.* lets each rule spend what it likes (held-out half, n=150; threshold tuned on the other half); *matched* equalises the reading budget on all cases, the comparison to read: it roughly halves the margin ($+0.05$–$+0.13$ on instruction-tuned models) and turns the apparent Llama tie into an efficiency win. All matched contrasts survive Holm adjustment ($p \leq 0.007$) except †MedQA (p=0.054). Per-setting CIs and full frontiers: Appendix B.4.

	rank		flip-stop	conf-stop	Δ (flip − conf)	
setting	(conf,cor)	ρ(flip)	acc @ reads	acc @ reads	self-term.	matched
RACE Qwen2.5-3B	0.32	0.51	0.893 @ 1.62	0.673 @ 0.35	+0.220	+0.130
RACE Qwen2.5-7B	0.42	0.49	0.927 @ 1.29	0.713 @ 0.37	+0.213	+0.050
RACE Qwen2.5-14B	0.36	0.64	0.940 @ 1.08	0.753 @ 0.25	+0.187	+0.093
RACE Qwen2.5-32B	0.38	0.62	0.960 @ 1.10	0.747 @ 0.29	+0.213	+0.070
RACE Llama-3.1-8B	0.52	0.36	0.827 @ 1.41	0.820 @ 2.91	+0.007	+0.097
MedQA Qwen2.5-7B	0.33	0.24	0.626 @ 1.63	0.483 @ 0.65	+0.143	+0.048†

5.3 The Flip Carries Stopping to LLM Beliefs

At matched reading budget flip-gated stopping beats confidence-gated stopping by $+0.05$–$+0.13$ on the instruction-tuned models (Table 1), and by $+0.24$/$+0.29$ on the two reasoning models. We report the matched contrast as the headline because the self-terminating margin ($+0.14$–$+0.22$) lets the flip rule read more; the matched number is roughly half of it, and MedQA does not survive Holm adjustment. What the matched comparison does not erode is the frontier claim: in all eight settings no point on the confidence frontier reaches the flip rule's accuracy at or below its budget, and confidence needs 3.5–5$\times$ the reads to catch up (Appendix B.4). On calibrated Llama the self-terminating tie ($+0.007$) is itself a budget artifact—matched, the flip rule wins $+0.097$—so the honest reading there is an *efficiency* advantage: it reaches the confidence frontier's ceiling with 1.45 reads against 5.14. The flip is not cheaper—confidence stops too early (overconfident), the flip stops when the answer stops changing, later but correct—and it also beats ER-gated stopping in 7/8 settings, both being *scale* quantities that collapse as the belief sharpens. A full calibrator family never raises the LLM magnitude ρ above ~ 0.12: the magnitude null is an information property, not a calibration artifact.

Reasoning models: the extreme of the miscalibration axis. Test-time compute is deployed on *reasoning* models, whose RL training drives final-letter logits near one-hot—where confidence is least informative. On two cross-family reasoning models (n=150 each) the within-step confidence–correctness rank is *inverted* (-0.326/-0.197), the magnitude saturates (98–99% of states $< 10^{-3}$), yet the flip retains rank information ($\rho = +0.464$/$+0.449$): flip-gated reading beats confidence stopping by $+0.24$/$+0.29$ at matched budget ($+0.387$/$+0.400$ self-terminating; $p < 2\times10^{-4}$) and recovers full-budget accuracy. Confidence is *inverted* on reasoning models, and confidence-gated stopping reads ≈ 0 findings to land at ≈ 0-info accuracy.

Across the eight settings the flip-over-confidence advantage is largest where the within-step confidence–correctness rank is inverted and near zero where it is highest (Fig. 8). We read the *self-terminating* margins there, because the statement is about how each rule behaves when left to choose its own budget; the relation is a *regime contrast* between rank-inverted and rank-positive beliefs and is not established within the instruction-tuned range alone (Appendix B.5).

5.4 One Currency, Three Actions: the Unified Controller

A single controller arbitrates **read/compute/stop** using decision value as one currency (Fig. 1C): reading scores the flip from one more finding, computing from one more sample, stop when neither carries positive decision value. Over a MedQA compute$\times$info grid (7B, n=196), Table 2:

The controller Pareto-dominates the deployed confidence-gated compute controller by $+0.255$ at lower cost (2.50 vs 3.92, $p < 2\times10^{-4}$), and no other threshold of it comes close: across its whole sweep (cost 0–3.92) confgate stays at 0.42–0.47, the zero-information floor, at least $+0.245$ below ours. It needs no stopping threshold and no reading-budget parameter (ε is degenerate by construction, τ_{starve} inert below 0.7). It also clears the *entire* pure-reading frontier—itself non-monotone, peaking at 0.673 (k=5) rather than at read-everything (0.663)—reaching 0.714 at 45% of that bud-

Table 2: The unified controller Pareto-dominates the deployed confidence-gated compute controller (every other confgate threshold, cost 0–3.92, stays at 0.42–0.47), and sits above the entire fixed-budget reading frontier—including its best point (0.673 at k=5), which already exceeds the read-everything endpoint (0.663).

controller	test cost	test acc	vs deployed confgate
ours (unified)	**2.50**	**0.714**	$+0.255$ $[0.153, 0.367]$, $p<2\times10^{-4}$
confgate (deployed)	3.92	0.459	—
confread	0.64	0.541	$+0.174$, p=0.0004
voi_fixed (oracle-tuned)	2.0–4.6	0.643–0.673	—

get. It takes 125 reads, 15 computes and 83 stops, admitting compute only at $L\geq1$: confgate's 3.92 samples *at zero information* reach 0.459, indistinguishable from the 0.449 floor, while the same compute after one or two findings is worth $+0.051$. Appendix C.4 gives the action mix, the knob sweeps, and why clearing the frontier is not leakage; the margin survives any $c_r/c_c \lesssim 2.94$.

What we could not beat. Four introspective sufficiency signals, 2-step non-myopic lookahead, an outcome-supervised logistic selector, three fully-trained acquisition policies and a learned stopper all halt at the same frontier below the gold-aware oracle, in both regimes (Appendix D). This is a tested frontier over the controllers we *implemented*, not an impossibility result over all model-output signals; the untested classes that would falsify us (ZIP-RC (Manvi et al., 2025), SVR (Chen et al., 2026)) are named.

6 Limitations

The LLM selection experiments score candidates on their *realized content*—a criterion scores candidate f by rendering the state already containing f's sentence (Appendix E). Every criterion is subject to this identically, so comparisons are fair, but the LLM criterion is a *realized entropy reduction*, not the *expected* information gain a Bayesian design computes over an observation model; we name it ER accordingly and reserve "EIG" for DDxPlus, where a Bayesian answer model makes it a genuine expectation (§5.2). Our LLM evidence is confined to 4-way multiple-choice decision spaces; the open-ended extension needs a semantic-equivalence judge with low paraphrase drift (§4.3), which we state as a gate and future work. **Compute baseline.** We compare against self-consistency and a grid-restricted gold-aware allocator of that budget, but do not implement verifier-guided search, process-reward models, or learned self-assessment (§3, §5); those are stronger baselines, and their absence is a limitation, not a claim of dominance. **Two-environment split.** The stopping result is two-tier by signal and environment: the magnitude signal on a synthetic classifier belief (its advantage a mid-budget regime, reversing at aggressive budgets), the flip signal on two interactive-LLM domains and two reasoning models—each environment carrying one layer of the thesis, which limits the combined claim. The flip's temperature invariance is constructive, backed by a single-domain check, not a full sweep. The three-factor ρ decomposition is a within-environment, cross-carrier attribution (illustrative rather than causal). The compute×info interaction is not resolved at our sample size; we rest on endpoint contrasts. Ethics statement in Appendix F.

7 Conclusion

Confidence-gated test-time compute rests on a signal that fails exactly when the model lacks the information to answer. Selecting on the *decision* rather than the distribution—training-free, gold-free, and distinct from the entropy-reduction objective the field uses—recovers accuracy the compute axis does not buy, across scales, a second domain and a second model family. Which layer pays off is set by *how* the belief is miscalibrated: ordering miscalibration rewards selection, level miscalibration rewards stopping, where the ordinal flip survives the saturation that destroys the continuous magnitude. The gap to a gold-aware oracle resists every realizable controller we tested: under insufficiency the fix is to read the right content, not to think harder about the wrong belief—within four-way multiple-choice reading over an available record and one controlled acquisition environment, not as a general replacement for confidence-based test-time control.

AI USE STATEMENT

The LLMs evaluated in this work—Qwen2.5-{3B,7B,14B,32B}-Instruct, Llama-3.1-8B-Instruct, and DeepSeek-R1-Distill-{Qwen-7B,Llama-8B}—are experimental *subjects*, not authoring tools; they are disclosed in §5 and Appendix E and are outside the scope of this statement. In this work, we used generative AI tools for pre-submission review that gave feedback on the experiments and on one mathematical claim; we did not use them to generate synthetic data, to propose the research hypotheses, for translation, or for dataset cleaning, and the remaining required-disclosure tasks are not applicable. Additionally, we used generative AI tools to edit the manuscript for readability. All AI-assisted output was reviewed against ground truth: every numerical claim in this paper is regenerated by the audit scripts from the logged run outputs, and AI-suggested wording and AI-generated code were checked against those same logs. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public benchmark datasets only—MedQA-USMLE (Jin et al., 2020) (CC-BY-4.0), DDxPlus (Tchango et al., 2022) (CC-BY-4.0), RACE (Lai et al., 2017), and MediQ (Li et al., 2024)—none of which contains protected health information; MIMIC-style data under a data-use agreement is deliberately excluded. No human subjects were involved, no personally identifying information was processed, and no new data were collected for this work. Although the testbed is clinical in surface form, the contribution is a decision-making signal for an agent under a compute budget, not a medical device: our claims concern test-time compute allocation, not clinical safety, and any clinician-facing deployment would require independent validation of the finding-selection policy against patient outcomes. We consider the main foreseeable misuse to be over-reading the MedQA results as evidence of diagnostic competence, and we scope the claims explicitly against that reading in §6. The full statement is Appendix F.

REPRODUCIBILITY STATEMENT

Every quantity in this paper is produced by a deterministic pipeline over released benchmarks, and the details needed to reproduce it are stated in the paper rather than deferred. Appendix E gives the full experimental protocol: the interactive-reading construction and the exact sentence-segmentation rule that defines the finding set, the model list and vLLM serving configuration, the verbatim system and user prompts with their decoding parameters, the belief-extraction and answer-parsing rules (including the final-token variant used for the reasoning models), the DDxPlus environment, and the failure convention for a dropped API call. The evaluation protocol is specified in the same appendix: threshold selection on a fixed 50/50 validation/test permutation seeded by `numpy.random.default_rng(0)`, paired bootstrap over cases with 5000 resamples, the resulting p-value resolution floor of 2×10^{-4}, and the two declared hypothesis families over which Holm–Bonferroni adjustment is applied. The formal claims (Propositions 1–4) are stated with their assumptions in §4.1 and proved in the appendix, and Claim 1 is decomposed in Appendix A. The negative controls that bound what we claim are enumerated in Appendix D.

Source code for the estimators, the interactive environments, and the plotting and audit scripts that regenerate every table and figure from logged run outputs is included in the supplementary material; the figure scripts run without a GPU from the logged data shipped alongside them.

REFERENCES

Linus Aronsson, Arman Rahbar, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. 2025. arXiv:2502.11067.

Raphaël Bentegeac, Bastien Le Guellec, Grégory Kuchcinski, Philippe Amouyel, and Aghiles Hamroun. Token probabilities to mitigate large language models overconfidence in answering medical questions: Quantitative study. *Journal of Medical Internet Research*, 27:e64348, 2025. doi: 10.2196/64348.

Ahsan Bilal, Ahmed Mohsin, Muhammad Umer, Ali Subhan, Hassan Rizwan, Ayesha Mohsin, and Dean Hougen. What if we allocate test-time compute adaptively? 2026. arXiv:2602.01070.

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

Hongyu Chen, Liang Lin, and Guangrun Wang. SVR: Self-verifying refinement via joint verdict-confidence reinforcement learning for adaptive test-time compute. 2026. arXiv:2607.28457.

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. Provable scaling laws for the test-time compute of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2411.19477.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. BED-LLM: Intelligent information gathering with LLMs and bayesian experimental design. 2025. arXiv:2508.21184.

C. Nicolò De Sabbata, Theodore R. Sumers, Badr AlKhamissi, Antoine Bosselut, and Thomas L. Griffiths. Rational metareasoning for large language models. 2024. arXiv:2410.05563.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human-agent communication. 2026. arXiv:2601.06407.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Joshua S. Gans. Optimal use of preferences in artificial intelligence algorithms. *NBER Working Paper 34780*, 2026. arXiv:2601.18732.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. 2024. arXiv:2407.21783.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Workshop on Trustworthy Natural Language Processing (TrustNLP)*, 2024. arXiv:2405.02917.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Jakob Hartmann, James Harvey, Jhonathan Navott, Erik Y. Wang, Luckeciano C. Melo, Flaviu Cipcigan, Cheng Zhang, and Alessandro Abate. Amortising bayesian experimental design for sequential information gathering in LLMs. 2026. FoGen @ ICML 2026 (non-archival). arXiv:2607.03426.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. 2020. arXiv:2009.13081.

Ardit Krasniqi, Luan Vejsiu, and Elira Dervishi. Granularity-regulated adaptive computational efficiency for optimal verification in test-time scaling. 2026. arXiv:2606.19354.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale reading comprehension dataset from examinations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. arXiv:1704.04683.

Shuyue Stella Li et al. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.00922.

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

Rohin Manvi et al. Zero-overhead introspection for adaptive test-time compute. 2025. arXiv:2512.01457.

John Ray B. Martinez. Multi-agent reasoning with consistency verification improves uncertainty calibration in medical mcqa. 2026. arXiv:2603.24481.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. 2020. arXiv:2006.01862.

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. 2025. arXiv:2501.19393.

J. Pablo Muñoz and Jinjie Yuan. RTTC: Reward-guided collaborative test-time compute. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24793–24809, 2025. arXiv:2508.10024.

Y. Peng, X. Zhu, C. Wei, N. Zeng, L. Wang, Y. T. He, and F. R. Yu. SAGE: Multi-agent self-evolution for LLM reasoning. 2026. arXiv:2603.15255.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Shuhui Qu. Adaptive test-time compute allocation via learned heuristics over categorical structure. 2026. arXiv:2602.03975.

Arman Rahbar, Ziyu Ye, Yuxin Chen, and Morteza Haghir Chehreghani. Efficient online decision tree learning with active feature acquisition. 2023. arXiv:2305.02093.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Safavi-Naini, Ali Soroush, and Jonathan H. Chen. Large language model uncertainty proxies: Discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149, 2025. doi: 10.1093/jamia/ocae254.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. 2024. arXiv:2408.03314.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. Structured uncertainty guided clarification for LLM agents. 2025. arXiv:2511.08798.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *NeurIPS Datasets and Benchmarks Track*, 2022. arXiv:2205.09148.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.14975.

Michael Valancius, Max Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. 2023. arXiv:2302.13960.

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. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiaxin Wu, Yizhou Yu, and Hong-Yu Zhou. Uncertainty estimation of large language models in medical question answering. 2024. arXiv:2407.08662.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13063.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. 2024. arXiv:2412.15115.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

Zhiyuan Zhai et al. Adaptive test-time compute allocation for reasoning LLMs via constrained policy optimization. 2026. arXiv:2604.14853.

Qiyuan Zhang et al. A survey on test-time scaling in large language models: What, how, where, and how well? 2025. arXiv:2503.24235.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. 2025. arXiv:2506.12721.

Table 3: Confidence–correctness correlation collapses under information insufficiency and recovers with more of the record. Confidence is self-consistency agreement—the majority-vote fraction over five chain-of-thought samples at temperature 0.7—and correctness is that of the single greedy (temperature 0) direct answer. Case bootstrap 2000, n=300.

model	minimal info	full info	gap
Qwen2.5-7B	$+0.023$ $[-0.09, +0.14]$	$+0.293$ $[+0.18, +0.40]$	$+0.27$
Qwen2.5-14B	$+0.085$ $[-0.03, +0.20]$	$+0.365$ $[+0.25, +0.47]$	$+0.28$
Qwen2.5-32B	$+0.147$ $[+0.03, +0.25]$	$+0.356$ $[+0.24, +0.48]$	$+0.21$

A Full Decomposition of Claim 1 (Entropy-Reduction ≈ Confidence Gain)

Claim 1 states that entropy-reduction selection and top-probability selection induce nearly the same ordering. It is not an identity on the simplex in general; we state and measure it as an empirical near-equivalence. Throughout, "ER" denotes greedy entropy-reduction selection as realized on our rendered-candidate testbed (the "EIG" rule of the literature); "EIG" is reserved for the DDxPlus environment, where a Bayesian answer model makes it a genuine expectation.

Geometry. For a belief $p \in \Delta^K$ write $m = \max_y p_y$, let $j = \arg\max_y p_y$, and let q be the residual distribution over the remaining $K-1$ classes, $q_i = p_i/(1-m)$ for $i \neq j$. Then

$$H(p) \;=\; H_b(m) \;+\; (1-m)\,H(q), \qquad H_b(m) = -m\log m - (1-m)\log(1-m).$$

Selecting to minimize posterior entropy is selecting to minimize $H_b(m') + (1-m')H(q')$; selecting to maximize confidence is selecting to maximize m'. The first term orders candidates exactly as confidence, and the second term—the only source of disagreement—is bounded by $(1-m')\log(K-1)$. As beliefs peak ($m' \to 1$), the disagreement term vanishes linearly in $1-m'$, so the two orderings coincide in the peaked regime. For K=2 they are exactly identical ($H(q) \equiv 0$). This predicts the correlation is weakest at flat (low-evidence) beliefs and rises as evidence accumulates—which is what we measure.

Measurement. On the DDxPlus trajectory (all $\approx 1.5\times10^5$ belief states visited by the acquisition policy), Spearman $\rho(-H, \max_y p_y) = 0.9922$. Across the six Qwen2.5 interactive-LLM runs, ER and VOI-conf select the *identical finding at every observed state* (selection-agreement Wilson lower bound ≥ 0.968); on Llama-3.1-8B agreement is 88.0% $[80.9, 92.7]$—the one divergence the geometry predicts, since Llama's beliefs are less peaked and the residual term has room to bite. The upshot: the SOTA finding-selection criterion (BED-LLM's greedy entropy reduction) is, on these belief distributions, operationally a confidence-family signal, and therefore inherits the §3 insufficiency pathology.

B Stopping: Full Tables and the Three-Factor ρ Decomposition

B.1 DDxPlus stopping on a shared acquisition trajectory

All rules run on the *identical* EIG acquisition trajectory; they differ only in the stopping signal, so per-patient correctness is paired (n=10^4, paired bootstrap 5000). The belief is order-calibrated (ECE 0.034, pooled over the test trajectory) but level-drifts: mean confidence$-$accuracy moves from -0.073 (0 evidence) to $+0.088$ (14 evidence).

The advantage is a mid-budget regime, bounded *before* reporting (we fixed the comparison windows ex ante, then ran once): decision-value stopping wins at the $\sim$5.5 and $\sim$7.0 matched budgets, attenuating as the budget grows toward reading everything, and reverses at $\sim$4.5, where it stops too early (Table 4). The temperature-control is decisive: recalibrating (temperature scaling halves ECE on the validation states it is fitted on, $0.026\to0.012$, Fig. 5, while on the test trajectory the same temperature raises it $0.034\to0.107$) does not rescue confidence-stopping, because (i) by Prop. 3(ii) monotone transforms of the stopping score leave the frontier pointwise invariant, and (ii) multiclass recalibration, the one transform that escapes that proof, reorders states in a way that moves the frontier the *wrong* way (-0.03 to -0.04). The scope of Prop. 3(ii) matters here: the invariance requires

Table 4: Decision-value stopping minus confidence- and EIG-stopping at matched acquisition budget. Every rule stops on the *same* EIG trajectory, so per-patient correctness is paired (n=10^4, paired bootstrap 5000; $^*p<2\times10^{-4}$). The advantage is a mid-budget one: largest near 5.5 acquisitions ($+0.0140$ $[+0.011, +0.017]$ over confidence), smaller at 7.0, reversed at 4.5 where decision-value stops too early. On the temperature-scaled belief (T=0.75) decision-value still beats confidence at every budget: recalibration does not close the gap (Prop. 3(ii)/(iii)).

	raw belief		temperature-scaled, T=0.75	
matched budget	vs confidence	vs EIG	vs confidence	vs EIG
∼4.5 acq	-0.018	-0.011	$+0.0154^*$	$+0.0104^*$
∼5.5 acq	$+0.0140^*$	$+0.0137^*$	$+0.0103^*$	$+0.0010$ (p=0.24)
∼7.0 acq	$+0.0079^*$	$+0.0048^*$	$+0.0035^*$	-0.0029^*

h *strictly* increasing and a function of $\mathrm{conf}(S)$ *alone*. Temperature scaling satisfies both, so it is inert; a multiclass recalibrator $\widetilde{\mathrm{conf}} = g(p_{\max})/\sum_j g(p_j)$ is a function of the full belief vector, so it escapes the invariance and can—and here does—move the frontier.

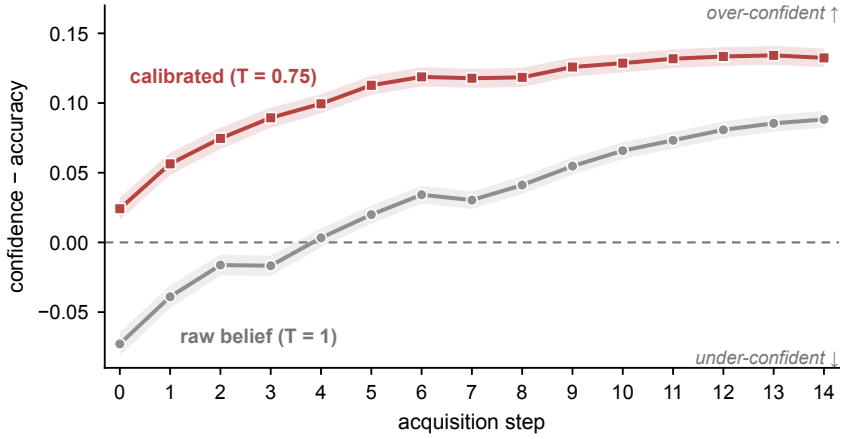


Figure 4: **The mechanism: the confidence *level* gap changes sign along the acquisition trajectory, and one temperature cannot fix both ends.** Mean confidence minus empirical accuracy at each acquisition step on the DDxPlus belief (n=10^4 patients; bands are 95% normal-approximation CIs). The raw belief (grey) is under-confident early (-0.073 at step 0), crosses zero between steps 3 and 5 (the step-4 CI contains 0), and is over-confident late ($+0.088$ by step 14). A single temperature is one monotone rescaling and cannot be correct at both ends: the NLL-fitted T=0.75 that halves ECE on its validation states (Fig. 5; red, EIG trajectory rerun on the calibrated belief) is over-confident at every step, by $+0.024$ to $+0.134$, so on this trajectory it raises pooled ECE from 0.034 to 0.107. Decision-value stopping reads a *marginal* quantity instead of a level and is invariant to this drift.

B.2 Three-factor decomposition of the DDxPlus→MedQA transfer gap

Why does decision-value stopping transfer to the DDxPlus classifier but the magnitude signal fail on MedQA? We decompose the transfer-gap in $\rho(\mathrm{signal}, \text{remaining-acc-gain})$ into three carriers: **granularity** (domain granularity of the evidence space), **sharpness** (peakedness of the belief, driving signal saturation), and **the residual**. The decomposition is a within-environment, cross-carrier attribution—illustrative, not causal. In the controlled environment granularity explains $\approx 39\%$ of the gap (Fig. 6); the residual is dominated by sharpness (the saturation of the continuous magnitude on peaked LLM beliefs, §5.2). The binary flip is the repair for exactly the sharpness term.

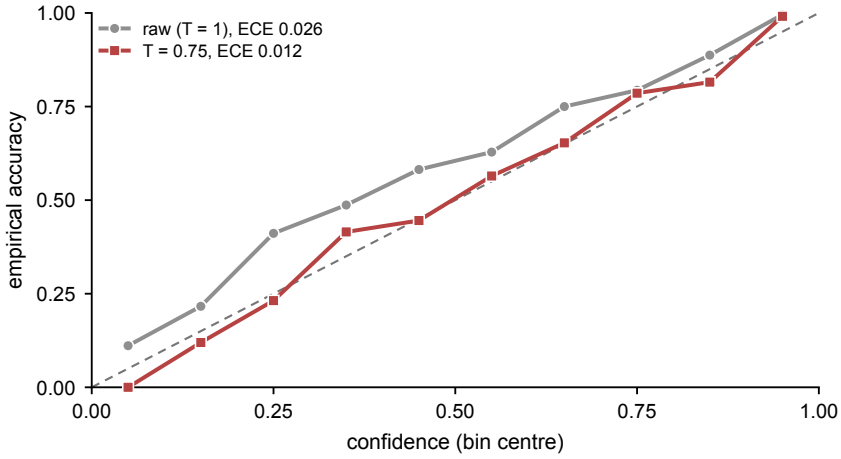


Figure 5: **Temperature scaling repairs aggregate calibration yet cannot repair stopping.** Reliability diagram for the DDxPlus partial-observation belief before (T=1, grey) and after (T=0.75, red; fitted by NLL on a validation split of 4 000 partial-observation states, each a random subset of a patient's evidence, disjoint from the test patients). The diagram and its ECE are computed on those same 4 000 states: pooled ECE falls $0.026 \to 0.012$ (15 bins; 10 plotted). Points are empirical accuracy per confidence bin, plotted at the bin centre; the dashed diagonal is perfect calibration. Halving ECE here does not help the stopper, because the stopper sees EIG-trajectory states, not random subsets: there the calibrated belief is over-confident at every step and its pooled ECE rises $0.034 \to 0.107$ (Fig. 4), and calibrated confidence stopping still loses to decision-value stopping at every budget ($p < 2 \times 10^{-4}$).

B.3 Flip-gated reading: the nulls that preceded it

The magnitude signal (continuous marginal decision value $\max_y p(y|S \cup f) - \max_y p(y|S)$) returned two nulls before we switched signals: MediQ (no reading headroom—acquisition does not move accuracy, so no stopping rule can help) and MedQA (headroom exists, but the magnitude $\rho \approx 0.086$). The latter is a signal property, not a domain property: $< 10^{-3}$ in 72–95% of states. A full calibrator family—temperature scaling, isotonic regression, and a non-monotone learned recalibrator over the features [confidence, entropy, margin, step index]—never raises the LLM magnitude ρ above ~ 0.12 and never lets magnitude-stopping beat calibrated confidence-stopping. This is the sharpest form of the thesis: the failure is in what the signal *is*, not in how it is tuned.

B.4 Cost-matched stopping frontiers on the LLM domains

Table 1 reports each rule at the budget it self-selects, and the flip rule reads more. That is not a matched comparison, so we build the accuracy-versus-budget frontier for every rule — the LLM analogue of Fig. 3 — by sweeping each threshold, and give the flip rule a budget knob by imposing a minimum number of reads (Fig. 7). Thresholds are swept on a grid that is *logarithmic near saturation* ($\tau = 1 - 10^{-1} \ldots 1 - 10^{-15}$, $\eta = 10^{-15} \ldots 1$): on peaked beliefs confidence sits at $1 - 10^{-9}$ and ER at 10^{-8}, and a linear grid truncates both baseline frontiers long before they reach "read everything," which would flatter our rule. Reproduce with `scripts/llm_stop_frontier.py`.

Two corrections to the main text follow from this table, and we state them rather than let the unmatched numbers stand. First, the headline stopping advantage on instruction-tuned models is $+0.05$–$+0.13$ at matched budget, not the $+0.14$–$+0.22$ obtained when each rule spends what it likes. Second, the reported "tie" on the well-calibrated Llama-3.1-8B ($+0.007$) is an artifact of budget mismatch — there confidence stopping read 2.91 findings against the flip rule's 1.41. At matched budget the flip rule wins on Llama too ($+0.097$, $p < 2 \times 10^{-4}$); what is true is that its advantage there is one of *efficiency*, reaching the confidence frontier's ceiling (0.840 at 5.14 reads) with 1.45 reads. Left to self-terminate, flip-gated reading stops at 1.0–1.6 findings and recovers full-budget accuracy at $\sim 4\times$ fewer reads (the star in Fig. 7).

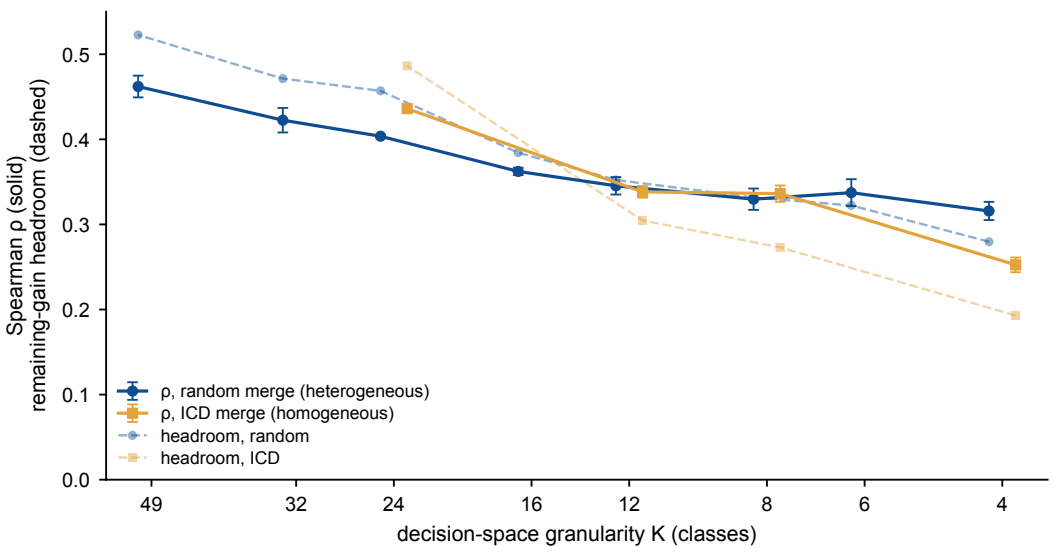


Figure 6: ρ **falls with decision-space granularity, and a clinically homogeneous merge does not rescue it.** ρ against the number of decision classes K on the calibrated DDxPlus environment (classifier retrained at every K, n=1000 held-out patients, 3 seeds; error bars are the across-seed s.d.). Solid curves are ρ; dashed curves are the remaining-gain headroom still available at that K. Both are dimensionless and share one axis, so the two merges can be compared directly on either quantity; K is on a log axis, and the two merges are drawn slightly apart in K so that their error bars do not overlap. Under a random balanced merge ρ falls from 0.462 at K=49 to 0.316 at K=4, monotonically except for an uptick at K=6 that is within seed error (0.330 at K=8, 0.337 at K=6); re-merging by ICD-10 chapter *deepens* rather than reverses the fall (K=4: ρ=0.252), because homogeneous groups are more separable and leave less remaining-gain headroom (0.193 vs 0.280 at K=4). Between K=12 and K=8 the two ρ curves are within seed error of each other (0.345 vs 0.338; 0.330 vs 0.336), so the separation between the merges is carried by the endpoints and not by the sweep's interior. ρ and headroom fall together across the whole sweep, which is why the three drivers cannot be read off K alone and must be attributed jointly.

Table 5: Flip- versus confidence-gated stopping *at matched reading budget*, on all cases of each setting. The confidence threshold is the one whose mean budget is nearest the flip rule's; [‡]where none lies within 0.1 reads (R1-Distill-Qwen-7B: no threshold stops between 1.91 and 5.8 reads) we mix the two bracketing thresholds so the expected budget matches exactly, scoring each case by its expected correctness under the mix. The advantage is roughly half what Table 1 reports at unmatched budgets, and MedQA does not survive Holm adjustment. In all eight settings no point on the confidence frontier reaches the flip rule's accuracy at or below its budget; confidence needs 3.5–5× the reads to catch up.

setting	flip acc @ reads	conf @ matched	Δ	95% CI	Holm p
RACE Qwen2.5-3B	0.873 @ 1.65	0.743 @ 1.70	+0.130	[+0.090, +0.173]	< 0.001
RACE Qwen2.5-7B	0.923 @ 1.27	0.873 @ 1.30	+0.050	[+0.017, +0.083]	0.007
RACE Qwen2.5-14B	0.943 @ 1.10	0.850 @ 1.08	+0.093	[+0.063, +0.127]	< 0.001
RACE Qwen2.5-32B	0.967 @ 1.01	0.897 @ 1.00	+0.070	[+0.040, +0.103]	< 0.001
RACE Llama-3.1-8B	0.837 @ 1.45	0.740 @ 1.50	+0.097	[+0.057, +0.137]	< 0.001
MedQA Qwen2.5-7B	0.592 @ 1.60	0.544 @ 1.62	+0.048	[+0.000, +0.095]	0.054
R1-Distill-Qwen-7B	0.793 @ 3.57	0.551 @ 3.57[‡]	+0.243	[+0.182, +0.305]	< 0.001
R1-Distill-Llama-8B	0.847 @ 3.03	0.553 @ 3.06	+0.293	[+0.207, +0.373]	< 0.001

B.5 Robustness of the miscalibration-axis relation

The advantage is smaller where the within-step confidence–correctness rank is higher, though not monotonically so within the rank-positive settings: Qwen2.5 on RACE (rank +0.32–0.42) gains +0.19–0.22, Qwen2.5-7B on MedQA (rank +0.33) +0.14, the better-calibrated Llama-3.1-8B

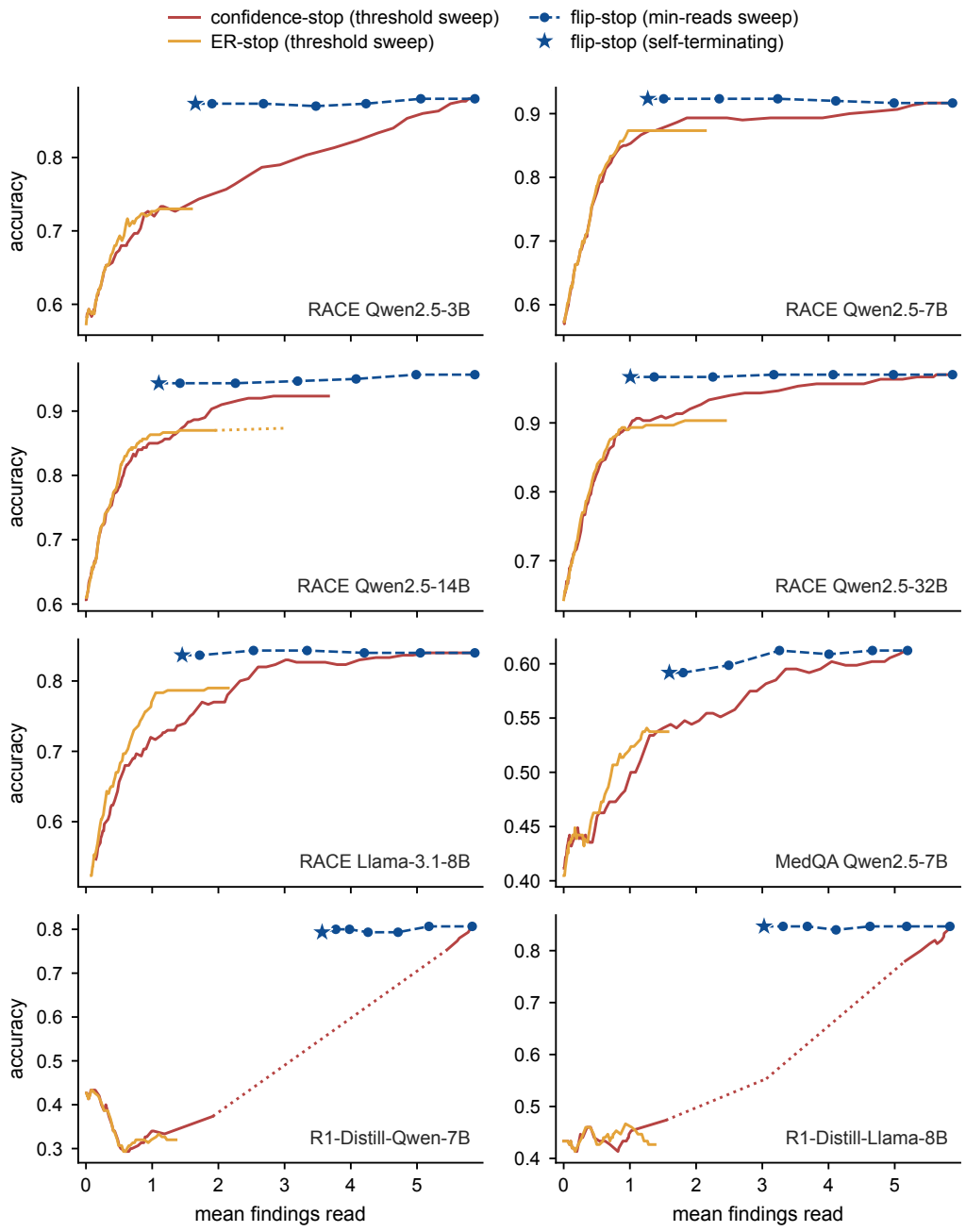


Figure 7: **Accuracy versus reading budget on all eight interactive-LLM settings** — the curves the DDxPlus figure (Fig. 3) shows for the controlled environment. Red and amber trace confidence- and ER-stopping as each rule's threshold is swept; a sweep is not a Pareto hull, so it can fall. Dotted segments join adjacent operating points more than one read apart, with nothing measured between them. Navy is the flip rule under a minimum-reads constraint, and the star is the flip rule left to self-terminate. From the star's budget upward the flip curve is at or above both baseline sweeps, and the three converge only when nearly every finding has been read. On the two reasoning models confidence-stopping accuracy falls over the first finding read. The y-axis range differs between panels.

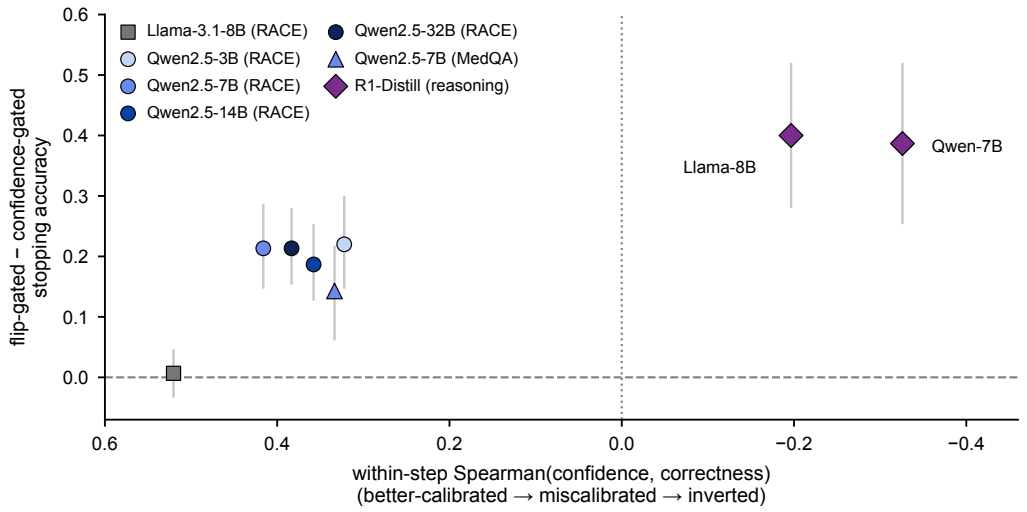


Figure 8: **The miscalibration axis.** Flip-gated minus confidence-gated stopping advantage (self-terminating, held-out half of each setting as in Table 1) against the within-step Spearman rank of confidence with correctness—both taken at the same step and pooled over every (question, step) state—one point per setting (paired-bootstrap 95% CI); colour is the model, marker shape the dataset, and the two reasoning points are labelled on the plot. Well-calibrated Llama-3.1-8B (rank $+0.52$) sits near zero, the instruction-tuned Qwen2.5 models (rank $+0.32$–0.42) win by $+0.14$–0.22, and the two reasoning models, whose near-one-hot beliefs *invert* the rank, win by the largest margins. The relation is carried by the contrast between the inverted and the rank-positive settings; it is not established within the instruction-tuned range alone (Appendix B.5).

(rank $+0.52$) gains $+0.007$, and the two rank-inverted reasoning models gain $+0.40$/$+0.39$ (Pearson -0.895, p=0.003; self-terminating margins, the convention appropriate to a statement about how each rule behaves when left to choose its own budget). Fig. 8 summarises eight heterogeneous, non-independent settings, so we test how much of the relation survives resampling and family removal (`scripts/miscal_axis_robustness.py`).

Table 6: The relation is carried by the contrast between rank-inverted (reasoning) and rank-positive (instruction-tuned) beliefs. Within the instruction-tuned models alone — a narrow rank range of 0.32–0.52 — it is not established.

subset	advantage as in Table 1		at matched budget	
	Pearson r	p	Pearson r	p
all eight settings	-0.895	0.003	-0.906	0.002
leave-one-out (worst case)	-0.847	—	-0.882	—
drop Llama-3.1 (1 setting)	-0.938	0.002	-0.933	0.002
drop both reasoning models	$\mathbf{-0.765}$	**0.076**	$\mathbf{-0.046}$	**0.93**

Leave-one-setting-out is stable (-0.85 to -0.94). Leave-one-*family*-out is not: with both reasoning models removed the relation is no longer significant under the paper's advantage definition ($r = -0.765$, $p = 0.076$, and dropping Llama from the remaining six flips the sign to $+0.393$), and it disappears entirely under the matched-budget definition ($r = -0.046$). We therefore claim a *regime contrast* — rank-inverted beliefs gain far more from the flip than rank-positive ones — and not a monotone law within the instruction-tuned range. A bootstrap over the eight points gives Pearson 95% CI $[-0.985, -0.543]$ for the reported advantage.

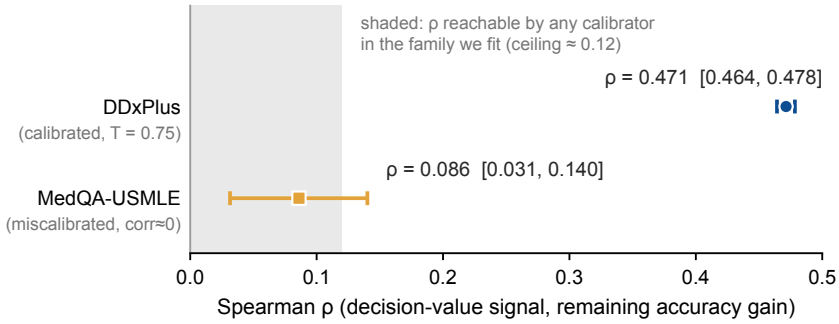


Figure 9: **The regime boundary is measured ex ante, not diagnosed post hoc.** Spearman correlation between the marginal decision-value signal and the *remaining accuracy gain* still available from revealing further findings, with 95% case-bootstrap intervals (2000 resamples; the resampling unit is the patient/question, not the case-step, since steps within a case are not independent). On the temperature-calibrated (T=0.75) DDxPlus belief ρ=0.471 $[0.464, 0.478]$ (n=10^4 patients), where decision-value stopping wins; on the miscalibrated MedQA-USMLE LLM belief (Qwen2.5-7B, temperature-softened) ρ=0.086 $[0.031, 0.140]$ (n=294 questions), where the magnitude signal saturates before accuracy plateaus. Prop. 3's corollary makes this ρ the ex ante *test* for transfer. The shaded band is the ρ any calibrator in the family we fit can reach (≈ 0.12): the LLM point estimate lies inside it and its interval ends just beyond it (0.140), far below the DDxPlus interval, so recalibration cannot move that environment across the boundary.

C Extended Selection Results

Per-budget accuracy at k=2 (the single-budget slice of Fig. 2A, retained here with the paired tests and the gold-aware ceiling):

model	FIXED	ER	**VOI-Δ**	Δ vs fixed [95% CI] (p)	ORACLE
7B	0.514	0.575	**0.644**	+0.130 [+0.048, +0.212] (0.0020)	0.829
14B	0.575	0.644	**0.705**	+0.130 [+0.055, +0.205] (0.0016)	0.863
32B	0.589	0.726	**0.753**	+0.164 [+0.089, +0.247] (0.0002)	0.890

The one-finding cold start. At k=1 VOI-Δ is below both ER and fixed order on the three Qwen2.5 scales (7B 0.411 vs 0.527/0.514; 14B 0.486 vs 0.596/0.527; 32B 0.521 vs 0.664/0.575) and ties fixed order on Llama-3.1-8B (0.487; ER 0.521). VOI-hybrid makes the first pick by confidence (the same pick as ER on the three Qwen2.5 runs) and continues by decision-VOI; it is at or above fixed order at every budget on all four models. From k=2 on VOI-Δ itself is at or above fixed order, random order and ER at every budget, on MedQA and on RACE.

Efficiency (findings-read-to-first-correct, lower better; MedQA, n=146):

model	FIXED	ER	VOI-Δ
7B	1.88	1.97	**1.43**
14B	1.81	1.79	**1.31**
32B	1.69	1.50	**1.03**

All VOI-Δ vs ER margins $p<2\times10^{-4}$; the paired-bootstrap 95% CIs for the ER−VOI-Δ difference are $[+0.33, +0.75]$ (7B), $[+0.29, +0.70]$ (14B) and $[+0.27, +0.67]$ (32B). A case whose answer is never correct counts as having read its whole record, which is why FIXED—the order with the most such cases—is not simply the largest entry in its column. Whole-curve AUC (findings 0–6) favors VOI-hybrid at every scale.

Llama-3.1-8B (second family, MedQA, n=117): divergence +0.161 $[0.135, 0.189]$, $p<2\times10^{-4}$—same magnitude as Qwen. **RACE** (n=150 per scale): mechanism correlation minimal→full rises +0.20 to +0.27; ER and VOI-conf agree on every observed selection; VOI-Δ most efficient (+0.23–0.28 over ER, $p<2\times10^{-4}$).

Domain-dependence, mechanism-predicted. Where minimal-info confidence is more valid (RACE option-guessing, Llama), VOI-Δ's *accuracy* margin at $k{=}2$ shrinks toward non-significance while its *efficiency* margin stays significant. This is not noise: Obs. 2 predicts decision-value selection pays off exactly where the belief *ordering* is wrong, and those datasets are less ordering-miscalibrated to begin with.

C.1 Cost on Both Axes (Probe-Free Baselines)

Every one-step-lookahead criterion (ER, VOI-conf, VOI-Δ) pays the identical probe bill (§4.3), so the headline VOI-Δ-vs-ER comparison is probe-matched and its margins are not bought with extra planning calls. Only FIXED and RANDOM are probe-free, so against them we report the margin on *both* axes: the $\sim$11 planning calls VOI-Δ spends per case buy $+0.13$–0.16 accuracy *and* 0.5–0.7 fewer findings read. VOI-Δ's probes need only a single greedy token, whereas ER needs the full distribution, so under a matched call count its calls are strictly the weaker requirement.

C.2 The Divergence, Measured Directly

At every sampled wrong-answer state (MedQA-7B, 700 paired states) we compare which finding each criterion picks:

Table 7: ER reduces entropy but sharpens a wrong belief; VOI-Δ targets the flip and corrects more often ($+0.167$, $p{<}2{\times}10^{-4}$).

criterion	P(flip)	P(correct | flip)	P(correct)
ER / BED-LLM	0.303	0.680	0.206
VOI-Δ	**0.537**	0.695	**0.373**

ER corrects only 20.6% of the time (it sharpens a wrong belief); VOI-Δ corrects 37.3%, $+0.167$ at identical reading cost ($p{<}2{\times}10^{-4}$; replicated at 14B: $+0.153$). The conditional rate is essentially criterion-independent (0.68 vs 0.70), so $+0.159$ of the $+0.167$ ($\sim$95%) is attributable to raising P(flip), and only $+0.008$ to choosing better among flipping candidates—the empirical content of Obs. 2, and the status of the tie-break.

C.3 The Gold-Aware Compute Allocator (Grid-Restricted)

Table 8: Compute cannot substitute for missing information (7B, MedQA, $n{=}117$).

info level	$m{=}1$	$m{=}4$	$m{=}8$	$m{=}16$
minimal (0 findings)	0.427	0.427	0.487	0.444
reveal 2 findings	0.624	0.675	0.692	0.718
full info	0.615	0.692	0.709	0.692

The per-case structure behind the allocator bound. Classifying every case by its behaviour across the *whole* compute sweep, 51 of 117 are wrong at every budget under minimal information, so no per-case allocator over this sweep—gold-aware or otherwise—can exceed 0.564 there, while revealing two findings scores 0.624 and cuts that band to 30. The band is read through its shrinkage across rows, not its absolute size; the cases that flip with m are $26/16/22\%$ across the three information levels, essentially flat, so that churn is sampling noise independent of information.

On the §5.4 grid ($n{=}196$) we compute a *grid-restricted gold-aware allocator* at minimal information: credit a case correct if *any* budget in the sweep answers correctly. We name it precisely, because it is an oracle only over what we swept—self-consistency at $m \in \{1, 4, 8\}$ samples, one sampling temperature (0.7), one prompt, one model. It upper-bounds *that* grid, not every compute-axis method and not larger budgets; we do not claim, and have not shown, that the winning candidate is absent at every conceivable budget. The grid-restricted allocator scores 0.505; revealing two findings with a single greedy pass scores 0.597—beating the oracle by $+0.092$ $[+0.005, +0.174]$, $p{=}0.044$, and the best fixed budget by $+0.138$, $p{=}0.001$. The gap is therefore not an artifact of a weakly *tuned*

compute baseline within this family; it does not rule out verifier-guided search, process-reward allocation, or learned self-assessment, none of which we implement (§6). Scaling-law guarantees (Chen et al., 2025) are driven by discriminating among sampled candidates; under insufficiency the winning candidate is absent from the samples at *any* budget, so the premise fails rather than the theorem.

C.4 Unified Controller: Action Mix and Budget

The controller self-stops at cost 2.50 with no stopping threshold and no reading-budget parameter, and sits marginally above the best *oracle-tuned* reading baseline (`voi_fixed`; $+0.071$ $[0.000, 0.143]$, p=0.056—we report this as marginal, not a win, since the interval touches zero) at 15–45% less budget than the `voi_fixed` settings nearest it (k=3, 4, 5 at cost 2.95, 3.81, 4.56). Its action mix on MedQA is 125 reads, 15 computes, 83 stops (1.28 reads/case, 0.153 computes/case): reading dominates, as §3 predicts, but compute is taken on a minority of cases and only at $L \geq 1$. That asymmetry is what lets the controller reach 0.714, above every point of the fixed-budget reading frontier: that frontier peaks at 0.673 (k=5, cost 4.56) and is non-monotone, since reading the last finding costs accuracy (k=6: 0.663 at cost 5.14). The controller beats its best point by $+0.041$ at 45% of the budget. **Why this is not leakage.** Exceeding the read-everything endpoint is available only by mixing the compute axis in, and the mechanism is mundane: on the cases where compute fires, the terminal answer is a self-consistency vote rather than a greedy decode, and voting is a different (here better) estimator of the same belief. No gold is consulted at any point—the flip test, the confidence gate and the vote are all observable—and the quantity cleared is the *greedy* full-information accuracy, not an oracle.

Knob sweeps. The "no tuned knob" claim is verified rather than asserted, and one knob is conceded. The flip threshold ε is degenerate by construction—both compared action values are binary, so every $\varepsilon \in (0, 1)$ induces the per-case identical rule across $\varepsilon \in \{0, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99\}$. The continuous fallback τ_{starve} is inert over most of its range: accuracy stays 0.714 for every value up to 0.9, while cost stays at 2.50 for $\tau_{\text{starve}} \leq 0.6$ and then climbs (2.60/2.61/2.63 at 0.7/0.8/0.9); at τ_{starve}=1 the fallback degenerates into "read everything" (cost 7.20, accuracy 0.643). It is inert below 0.7 because the first-token confidence on this dump is saturated (mean 0.972; only 3.4% of states below 0.7)—the §3 mechanism reappearing inside the controller. Between τ_{starve}=0 and 1, 19 of 98 cases change their final answer, so the flatness is a statement about the sweep, not about individual cases. *The compute-probe size m_{probe} is a genuine hyperparameter and we do not claim otherwise*: $m_{\text{probe}} \in \{1, 4, 8\}$ gives cost 1.44/1.85/2.50 and accuracy 0.663/0.704/0.714, and we select it on validation like any budget knob. The *reading* behaviour is invariant across that sweep (125 read actions at every m_{probe}); what changes is how much voting is bought once reading stops paying.

D Negative Controls: the Realizable Frontier

We implemented and tested the following as negative controls. None beats the myopic decision-VOI estimators; all halt at the same frontier, well below the gold-aware oracle, in both regimes (all $n \geq 10^4$ on DDxPlus, leakage-checked).

- **Introspective sufficiency signals (4):** four signals derived from the model's own distribution (margin, entropy, agreement, and a learned combination) used to gate reading. None improves over confidence-stopping; each is a confidence-family quantity by Prop. 3.
- **2-step non-myopic lookahead:** selecting by the decision value two findings ahead, not one. Improves nothing over one-step (the marginal decision value of the second step is captured by the first when beliefs are peaked), at $\sim n_f^2$ probe cost.
- **Outcome-supervised learned selector (LLM):** a logistic selector trained on held-out states to predict which finding would correct the answer. Fails to beat VOI-Δ on test states—it overfits the train belief distribution, and the signal it learns is again a confidence-family quantity.
- **Three fully-trained acquisition policies (DDxPlus):** an outcome-supervised policy, an EIG-reward policy, and a decision-value-reward policy, each trained to convergence. The

decision-value-reward policy is the best of the three and still does not beat the myopic VOI-Δ estimator, consistent with the amortization-gap observations in (Hartmann et al., 2026).

- **Learned stopper (DDxPlus):** a stopping policy trained on the same trajectory. Matches, but does not beat, decision-value stopping at matched budget.

Scope of the claim. This is a tested frontier over the controllers we *implemented*, not an impossibility result over all model-output signals. The untested classes that would falsify us are signals *trained into* the model's representation rather than read off its output distribution—ZIP-RC (Manvi et al., 2025) and SVR (Chen et al., 2026)—which we name rather than dismiss.

E Experimental Protocol

Interactive reading construction. For MedQA-USMLE and RACE, each vignette/passage is split into per-sentence *findings*; the agent observes the question, the options, and a revealed subset S of findings. "Minimal information" reveals none; "full information" reveals all. A criterion scores candidate finding f by rendering the state that already contains f's sentence (one greedy token) and evaluating the resulting belief. *Every* compared criterion is subject to this identically, so all comparisons are fair; the consequence is that §5.1 measures selection *among available-but-unread findings* (an ordering over an existing record), while the DDxPlus environment measures genuine acquisition. A direct consequence is a naming distinction we keep throughout: because the candidate's content is rendered into the state before the score is read, the LLM criterion realizes a *reduction of entropy after observing the content*, not the *expected* information gain that a Bayesian experimental design computes by averaging over an observation model. We therefore call the LLM criterion entropy reduction (ER) and reserve "EIG" for the DDxPlus environment, where a Bayesian answer model supplies the observation model and the quantity is a genuine expectation. The numbers are unchanged; only the name of the quantity is honest.

Models and serving. Qwen2.5-3B,7B,14B,32B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and two reasoning models (DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B). All served with vLLM (Kwon et al., 2023), temperature 0 for greedy probes, temperature 0.7 for self-consistency samples. Confidence estimated two ways (self-consistency agreement and first-token log-probability) to rule out proxy artifacts; the mechanism holds for both.

Sentence segmentation and the finding set. Findings are produced by splitting on `re.split(r'(?<=[.?!])\s+', text)`; for RACE newlines are replaced by spaces first. A case is skipped if it yields fewer than two sentences. The finding budget is capped at $n_f \leq 6$: if a passage has more than six sentences the first five are kept and the entire tail is merged into the sixth, so no text is discarded. For MedQA the *question* sentence (the last sentence containing "?", else the final sentence) is always shown and is excluded from the finding set; for RACE the question is always shown and all passage sentences are findings. The state is rendered by concatenating the revealed findings *in original passage order* (not in reveal order), so two different reading orders that reach the same set render identically.

Prompts and decoding. The probe is a single chat completion. System message: "Answer this USMLE question. Give ONLY the letter (A/B/C/D)." (MedQA) or "Answer the reading-comprehension question. Give ONLY the letter (A/B/C/D)." (RACE). User message: the rendered state, then `\nOptions:\n` and the four `A. - D.` option lines. Decoding is `temperature=0`, `max_tokens=4`, `logprobs=True`, `top_logprobs=20`. The reasoning models use the system message "You are a helpful assistant.", append "Think step by step, then give the final answer as a single letter (A/B/C/D)." to the user message, and `max_tokens=512`.

Belief and answer extraction. The belief is a softmax over the *first generated token's* top-20 logprobs restricted to $\{A, B, C, D\}$ and renormalized, at belief temperature T (T=1.0 throughout unless stated; T=2.5 is the softened MedQA variant). For the reasoning models the belief is read at the *final* letter token rather than the first, because a chain of thought mentions every letter and a first-token belief degrades to near-uniform. The answer letter is parsed as `answer[^A-D]*([ABCD])` on the uppercased completion, else the last standalone `[ABCD]`, else the argmax of the letter distribution. A failed API call yields a uniform belief and is counted as incorrect.

Threshold selection. For the stopping tables, cases are split 50/50 into validation and test by a fixed permutation (`numpy.random.default_rng(0)`); the confidence, EIG and magnitude thresholds are chosen on validation to *maximize accuracy*, which is generous to the baselines, and are then applied unchanged to test. The flip rule has no threshold to tune. For the cost-matched frontiers (Appendix B.4) no split is needed, because the baseline threshold is chosen to match a *budget* and never sees a label.

Sampling and significance. Reported n is the *analysed* sample after the skip rule below, not the number of items drawn. MedQA selection: 150 drawn / 146 analysed for the three Qwen2.5 scales and $120/117$ for Llama-3.1-8B (cases yielding fewer than two sentences are skipped). n=117 (compute grid), n=196 (controller grid), n=150/scale (RACE, no case skipped; reasoning models), n=10^4 (DDxPlus stopping). All hypothesis-test comparisons use a paired bootstrap with 5000 resamples, two-sided, reported with confidence intervals; the correlation intervals of Table 3 use 2000 resamples and are descriptive, so the 2×10^{-4} floor below does not apply to them; report both the finding-budget and total-LLM-call axes where probe cost differs. The resampling unit is the *case* (one question with its whole trajectory), never the step, so within-case dependence is respected. The two-sided p is read off the bootstrap distribution, so its resolution floor is $1/5000 = 2 \times 10^{-4}$: we therefore report the floor itself rather than a smaller number, and every "p<2×10^{-4}" in this paper means "below the bootstrap resolution", not an estimated p of that magnitude.

Multiplicity. We report raw p alongside Holm–Bonferroni-adjusted p *within* a declared family of hypotheses, rather than across the whole paper. The stopping family is the eight flip-versus-confidence comparisons of §5.2: at matched budget seven of the eight survive Holm adjustment, and MedQA does not (raw p=0.054, adjusted p=0.054) — see Appendix B.4. The selection family is the three scale comparisons of Fig. 2A (raw $p = 0.0008, 0.0008, <0.0001$; all Holm-adjusted $p < 0.01$). The remaining tested contrasts are declared *a priori* as one prespecified primary contrast plus secondary contrasts, and are reported at nominal p without adjustment. In §3 the prespecified primary contrast is reveal-two-findings versus $16\times$ compute at minimal information (p=0.0004); compute-at-minimal (p=0.72), compute-at-full (p=0.040) and the comparison against the grid-restricted gold-aware allocator (p=0.044) are secondary. We state this explicitly because Holm adjustment *within* a four-member family would retain only the primary contrast (adjusted p=0.0016) and would not retain the two contrasts near α (adjusted p=0.12 and 0.088); readers who prefer that convention should read those two as suggestive rather than confirmatory. Descriptive quantities reported without a hypothesis test (correlations, saturation rates, budget curves) are not part of any family.

DDxPlus environment. 49 pathologies, 223 evidences. A partial-observation classifier supplies the belief $p(\mathrm{dx} \mid \text{evidence})$; a Bayesian answer model makes EIG a genuine expected-information-gain competitor. Acquisition raises accuracy $0.06 \rightarrow 0.997$. Rules run on the same EIG trajectory and differ only in the stopping signal (paired).

F Ethics Statement

This work uses public benchmark datasets only: MedQA-USMLE (Jin et al., 2020) (CC-BY-4.0), DDxPlus (Tchango et al., 2022) (CC-BY-4.0), RACE (Lai et al., 2017), and MediQ (Li et al., 2024). None contains protected health information (MIMIC-style DUA data is deliberately excluded from the repository). The method is a decision-making signal for an agent under budget, not a medical device; its claims are about test-time compute allocation, not clinical safety, and a clinician-facing deployment would require independent validation of the finding-selection policy against patient outcomes. No human subjects, no personally identifying information, no data collected for this work.