COMPUTATION BEFORE CORRECTION: SEPARATING DIAGNOSTIC AND SOLUTION VALUE IN LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time reasoning is shifting from simply scaling computation to deciding when and where additional computation is worth spending. We ask whether reasoning can be useful before it can solve the problem: can short computation reveal whether substantially more, independently restarted reasoning is likely to succeed? We study this as *diagnostic computation* using a *probe-and-discard* design that separates solution value from observer-relative diagnostic value. From a frozen wrong reasoning state, a short probe is generated, observed, and discarded; evaluated long repairs restart independently from the original pre-probe state, preventing probe tokens from contributing to the repair itself.

In a prospectively frozen Qwen3-8B/MATH experiment, a 16-token probe directly corrects only 1.50% of cases yet substantially improves prediction of future repairability. A separately frozen DeepSeek/GSM8K experiment reproduces the qualitative separation across model family and benchmark. More broadly, the results suggest that test-time computation can play two roles: *advancing a solution or measuring whether further computation is worth spending*. This motivates treating solution value, diagnostic value, and decision value as distinct quantities rather than assuming that useful computation must immediately improve the answer.

1 INTRODUCTION

Inference-time reasoning is becoming a resource-allocation problem. A system that can spend additional computation on a failed state must still decide which states are worth that expenditure. Maximal reasoning wastes computation on failures that may remain unrecoverable, while stopping too early abandons states that additional reasoning could repair. The challenge is therefore not only whether more computation helps, but where it is likely to pay off. Recent work addresses this through adaptive allocation, return prediction, trajectory monitoring, and early exit (Snell et al., 2025; Damani et al., 2025; Manvi et al., 2024; 2026; Zabounidis et al., 2025; Zhou et al., 2026; Yang et al., 2026; Wang et al., 2026).

Most approaches treat inference-time computation primarily as an *action*: computation is valuable when it advances a solution. Self-consistency and repeated sampling likewise use extra trajectories to improve solution search (Wang et al., 2023; Brown et al., 2024). We study a second possibility. *Useful computation may have value before it improves the answer, because it can reveal whether more computation is worth performing.* If a short reasoning sample distinguishes failures that remain repairable from those that do not, reasoning becomes an *active computational measurement* for selective inference. This may matter in code repair, scientific reasoning, planning, or agentic workflows where systems decide whether to escalate uncertain states to more expensive reasoning. We do not claim deployment savings here; we ask whether computation can acquire predictive value before substantial solution value. Figure 1 positions this role and previews the primary separation.

The distinction is easy to confound. Failed reasoning states can differ sharply in *repairability*: some may recover with substantially more computation, while others may almost never recover. Yet if a short continuation is retained and reasoning continues from it, any improvement may simply reflect solution progress carried forward by those tokens. To determine whether short computation is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

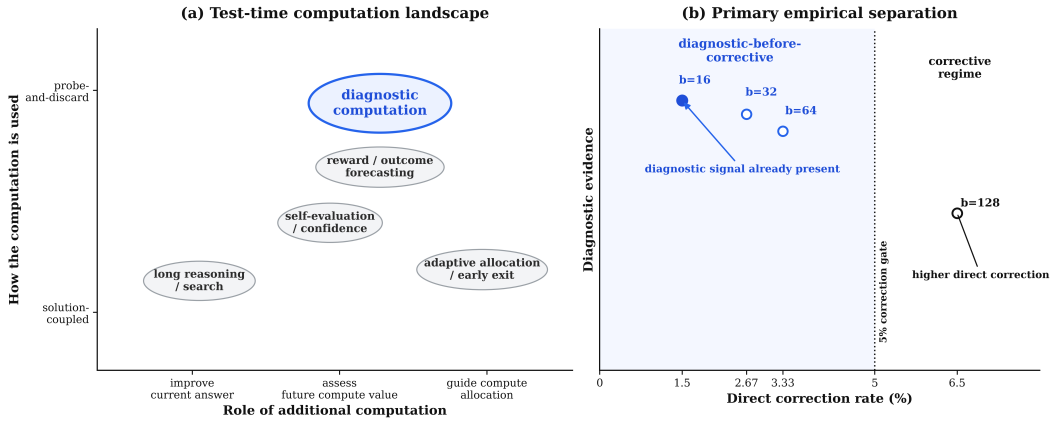

Figure 1: **Diagnostic computation in the test-time-compute landscape.** (a) Schematic positioning relative to nearby test-time-compute paradigms. (b) Schematic summary of the prospectively frozen primary budgets. Horizontal positions are the observed direct-correction rates; vertical placement summarizes diagnostic evidence qualitatively rather than encoding an effect-size scale. Diagnostic evidence appears while direct correction remains small.

informative rather than merely helpful, it must be observed without being allowed to assist the future solution.

We therefore develop a *probe-and-discard* design. From a frozen failed state, a short ordinary-reasoning branch is generated, observed, and discarded. The evaluated long-repair branch independently restarts from the original pre-probe state, so probe tokens cannot enter the evaluated repair trajectory. If observing the discarded probe improves prediction of whether the independent repair will succeed, the gain is diagnostic beyond the prespecified observer. Figure 2 summarizes this construction.

This setup exposes three kinds of value. A short computation may directly fix the answer. It may fail to fix the answer yet still reveal whether more reasoning is likely to succeed. Even useful diagnostic information may not justify its cost in a downstream policy. We call these *solution value*, *diagnostic value*, and *decision value*:

$$\boxed{\text{solution value} \neq \text{diagnostic value} \neq \text{decision value.}} \tag{1}$$

Solution value concerns immediate correction; diagnostic value concerns improved prediction of an independently restarted repair; decision value concerns whether that information improves compute allocation after costs are counted. Classical metareasoning studies computation through later decisions (Russell & Wefald, 1991; Hay et al., 2012); our goal is to separate these quantities empirically in LLM reasoning.

Our question is distinct from showing that reasoning contains information about correctness in general. Prior work shows that hidden states, confidence, early representations, and reasoning steps can predict correctness, failure, or future reward (Zhang et al., 2025a; Yoon et al., 2025; Zhang et al., 2025b; Ton et al., 2025; Qian et al., 2025; Bigelow et al., 2025; Bogdan et al., 2025); other methods forecast value from inputs or retained prefixes (Damani et al., 2025; Manvi et al., 2024; 2026; Zabounidis et al., 2025; Chen et al., 2025). Our target is narrower: whether a *discarded sample of ordinary reasoning* adds information about a separately sampled future-repair intervention.

Contributions. We make three contributions. First, we define observer-relative diagnostic computation and introduce a branch-separated protocol that distinguishes computation-for-measurement from computation-for-solving. Second, a prospectively frozen Qwen3-8B/MATH experiment establishes diagnostic-before-corrective separation, and a separately frozen DeepSeek/GSM8K experiment reproduces the qualitative effect across model family and benchmark; later cohorts provide boundary and robustness evidence rather than a universal token rule. Third, we test difficulty controls and a separately frozen stronger-observer sensitivity, characterize regime dependence, and evaluate

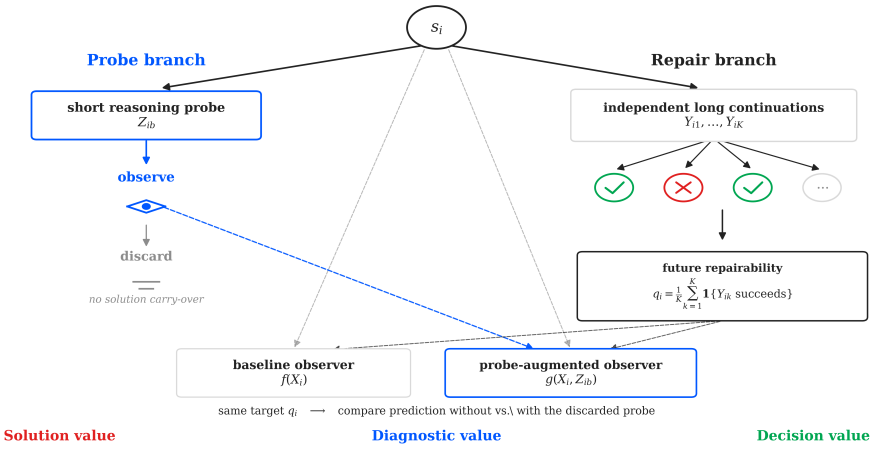

Figure 2: **Probe-and-discard identification of diagnostic computation.** From the same frozen failed state S, diagnostic and repair branches are independently generated. The short probe is observed and discarded; future repairability is measured by independent long continuations restarted from the pre-probe state. Comparing observers with and without the discarded probe isolates diagnostic value, while downstream allocation defines decision value.

compute-matched allocation. The resulting picture is two-sided: short reasoning can carry diagnostic information before substantial correction, but diagnostic information alone does not guarantee positive policy value. More broadly, the work motivates a *measure–decide–act* view of adaptive inference in which computation can help determine whether further computation should be purchased.

2 Diagnostic Computation: Problem and Identification

Intuitively, one failed state produces two separate futures: a short diagnostic branch and an independently restarted long repair branch, with no token carry-over.

Let S denote the complete frozen generating state of a baseline-wrong example. Let $O = \phi(S)$ denote the information available to a prespecified diagnostic observer before probing. A short probe Z_b of budget b and a binary repair outcome $Y \in \{0, 1\}$ from an independently sampled long continuation are generated conditional on S:

$$Z_b \perp Y \mid S. \tag{2}$$

Thus the probe cannot help by contributing tokens to the evaluated repair, but it may still reveal information about the shared frozen state.

What branch separation identifies. Equation (2) removes one confound: probe tokens cannot become part of the repaired solution. It does *not* make the probe exogenous to S; both branches still arise from the same frozen state. The probe can therefore remain predictive when O misses repair-relevant structure in S. Our estimand is this incremental predictive value for a bounded observer, not a causal effect of probe tokens on repair success.

Diagnostic value. Diagnostic value asks whether seeing the discarded probe improves repair-success prediction. Formally, define $\eta_0(O) = \Pr(Y = 1 \mid O)$ and $\eta_b(O, Z_b) = \Pr(Y = 1 \mid O, Z_b)$. Under Brier loss (Brier, 1950), the Bayes value of observing the probe relative to O is

$$D_b^{\mathrm{Br}}(O) = \mathbb{E}[(Y - \eta_0)^2] - \mathbb{E}[(Y - \eta_b)^2] \tag{3}$$

$$= \mathbb{E}[(\eta_b - \eta_0)^2] \geq 0. \tag{4}$$

Under log loss the corresponding quantity is $I(Y; Z_b \mid O)$. These are *observer-relative*. If O were the complete state S, Eq. (2) would imply zero incremental Bayes information. The empirical question is whether the short probe reveals repair-relevant structure not captured by the frozen observer.

This nonnegativity is a population identity for Bayes predictors; empirical $\Delta\mathrm{Br}$ from frozen finite-sample observers may be positive or negative.

Solution value. Let C_b indicate whether the short probe branch itself produces a correct terminal answer under the frozen scorer. Its direct solution value is $V_{\mathrm{sol}}(b) = \Pr(C_b = 1)$. This endpoint criterion does not assert that a short branch makes no intermediate progress; it asks only whether the branch supports a correct terminal answer at the scoring point. We are interested in budgets where diagnostic gain is positive while this rate remains below a prespecified small-correction gate.

Decision value. A useful signal is not automatically a useful policy. If a long continuation has value v, cost c_L, and the probe costs c_b, then the gross Bayes value of diagnostic information is

$$V_b^{\mathrm{dec}} = \mathbb{E}[(v\eta_b - c_L)_+] - \mathbb{E}[(v\eta_0 - c_L)_+], \tag{5}$$

and the net value of diagnostic computation is $\mathrm{NVDC}_b = V_b^{\mathrm{dec}} - c_b$. Positive diagnostic value therefore need not imply positive realized policy value for a learned finite-sample controller. Formal statements and proofs are in Appendix B.

3 Experimental Design

Why mathematical reasoning? We use mathematical reasoning as a controlled setting for identifying the phenomenon, not as the claimed scope of diagnostic computation. Contest-style problems are useful because they require multi-step reasoning, failed states can differ sharply in later recoverability, and final answers can be scored comparatively cleanly without subjective judges or learned preference models. This makes it possible to ask a focused question: does short computation *advance* the solution, or does it instead *measure* whether substantially more computation is likely to help?

The underlying allocation problem is not specific to mathematics. Similar decisions arise in code repair, scientific reasoning, planning, and agentic workflows whenever an uncertain or failed state may or may not justify further search, revision, reflection, or repair. We therefore use mathematics as a controlled testbed, not as evidence of automatic cross-domain transfer. Section 4 takes one step beyond the primary setting by testing the same qualitative separation with a different model family and benchmark. Broader generalization to coding, scientific reasoning, planning, and agentic tasks is not established here and remains an empirical question.

Primary model, population, and repair target. The primary experiment uses Qwen3-8B (Yang et al., 2025) on MATH (Hendrycks et al., 2021), an open-weight reasoning setting that permits repeated branch sampling under a frozen protocol. The target population is intentionally conditional on failure: we prospectively freeze 200 states whose baseline reasoning attempt is wrong. This matches the decision problem of interest—once an error has occurred, is substantially more computation worth attempting?—rather than benchmark-wide correctness prediction. For each state i, we independently generate $K = 20$ repair continuations of at most 512 tokens from the same pre-probe state and define

$$\widehat{q}_{20,i} = \frac{1}{20}\sum_{k=1}^{20} Y_{ik}. \tag{6}$$

Across the 200 states, 1,258 of 4,000 repair continuations succeed, giving mean $\widehat{q}_{20} = 0.3145$. Figure 3 shows the resulting heterogeneity across frozen wrong states, which defines the diagnostic prediction target. Throughout the paper, *repairability* is intervention-specific: it denotes success under this fixed model, continuation policy, scoring rule, and 512-token horizon, not intrinsic mathematical difficulty. Our estimand is therefore conditional repair behavior on frozen failed states, not benchmark-level generalization accuracy.

Finite-K target uncertainty. Repairability is estimated from only $K = 20$ future repairs per state, so $\widehat{q}_{20,i}$ is treated as a finite-sample measurement rather than a noise-free latent probability. The confirmatory target is the out-of-fold predictive risk for this same frozen measurement process applied to every state. Uncertainty analyses therefore resample whole states together with all 20 attached repair outcomes, rather than treating individual rollouts as independent sampling units. We report repair-trial Brier risk. Representing those 20 outcomes individually or through their empirical success rate changes the absolute loss only by a prediction-independent term, so the observer-to-observer difference $\Delta\mathrm{Br}$ is exactly the same in either representation; Appendix D gives the identity.

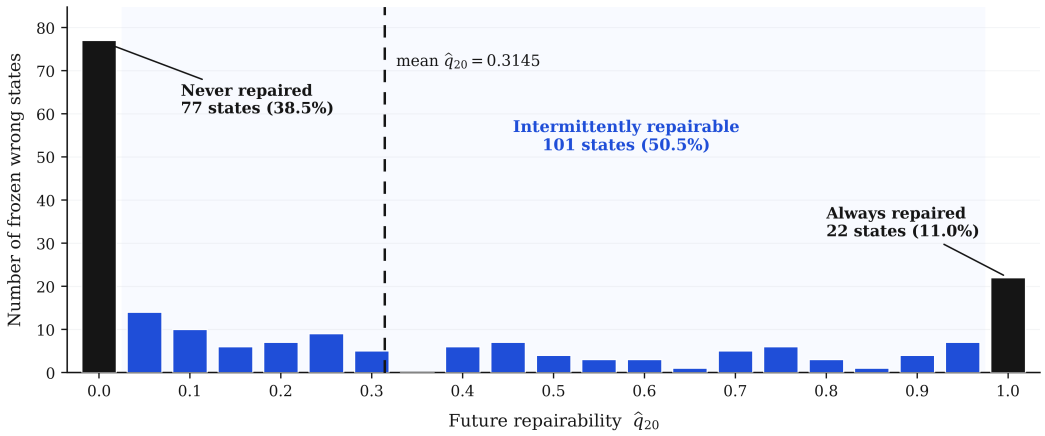

Figure 3: **Future repairability is highly heterogeneous across frozen wrong states.** Distribution of $\widehat{q}_{20}$ over 200 baseline-wrong Qwen3-8B/MATH states, estimated from 20 independent 512-token repair continuations per state. Seventy-seven states (38.5%) are never repaired, 101 (50.5%) are repaired intermittently, and 22 (11.0%) are repaired in every trial; mean repairability is 0.3145. This heterogeneity defines the diagnostic prediction target.

Diagnostic probes. For each frozen state, we independently sample three ordinary-reasoning probes at $b \in \{16, 32, 64, 128\}$, yielding 2,400 probe samples. Probes and long repairs use the same locked revision intervention—problem, frozen failed trace, and current answer—and the same sampling law ($T = 0.7$, top-$p = 0.9$); only branch randomness and the token cap differ. Each probe is observed and discarded and is never exposed to an evaluated repair continuation. The deterministic answer readout is scoring-only and does not feed tokens back into either branch. Exact serialization, model revisions, seed namespaces, and endpoint-scoring rules are in Appendix C.

Frozen observer and diagnostic estimand. The prediction task is deliberately simple: compare the same fixed observer before and after adding the discarded probe.

Five state-level folds are fixed before diagnostic evaluation. The baseline observer receives the problem, failed reasoning trace, and failed answer; the augmented observer additionally receives the discarded probe. The primary observer deliberately uses a fixed, lightweight word/character TF–IDF representation with ℓ_2-regularized logistic regression and no diagnostic-outcome-based model selection. It is not intended to exhaust all information in the state: diagnostic value is explicitly observer-relative. Section 4 therefore tests a stronger static observer and additional difficulty controls. For the primary Qwen pooled estimate, one augmented model per fold is fit jointly to the three probe rows per state; each repeated row carries one third of the state's binomial training weight, and the three held-out probe probabilities are averaged to one state-level prediction. Separate replicate-specific fits are used only for the 2-of-3 consistency screen. The historical DeepSeek replication retains its originally frozen joint-pooling specification; Appendix D gives the exact weighting and aggregation rules. We report

$$\Delta\mathrm{Br}(b) = \mathrm{Br}(\text{baseline}) - \mathrm{Br}(\text{baseline} + Z_b), \tag{7}$$

Thus $\Delta\mathrm{Br}(b) > 0$ has a simple interpretation: after holding the observer and evaluation protocol fixed, adding the discarded probe improves prediction of future repairability.

Prospectively specified gate. We call a budget a successful *diagnostic-before-corrective separation* only when the probe improves prediction reproducibly while direct correction remains rare. Concretely, all four prespecified conditions must hold: (i) pooled $\Delta\mathrm{Br} > 0$; (ii) a 10,000-draw fold-preserving full-pipeline probe-reassignment test is significant after Holm correction over $b \in \{16, 32, 64\}$; (iii) at least two of three replicate-specific gains are positive; and (iv) direct solution value is at most 5%. The 2-of-3 rule is a consistency screen rather than a second significance test, and the reassignment test is a state–probe pairing robustness check rather than an exact test of conditional independence given the baseline observer. The 128-token condition is descriptive because it lies outside the primary family and exceeds the correction gate. The 5% threshold is a prospectively

Table 1: Prospectively frozen primary design. Splits, budgets, observer choices, and gates were fixed before evaluating diagnostic outcomes on the valid primary population.

Component	Frozen specification
Population	200 baseline-wrong Qwen3-8B/MATH states
Repair target	20 independent continuations/state, 512-token maximum
Probe budgets	16, 32, 64 primary; 128 descriptive
Probe replicates	3 independent probes/state/budget
Observer evaluation	5 frozen state-level folds; out-of-fold Brier risk
Inference	10k state bootstraps; 10k full-pipeline reassignments
Primary gate	$\Delta\mathrm{Br} > 0$, Holm test, $\geq 2/3$ reps, correction $\leq 5\%$

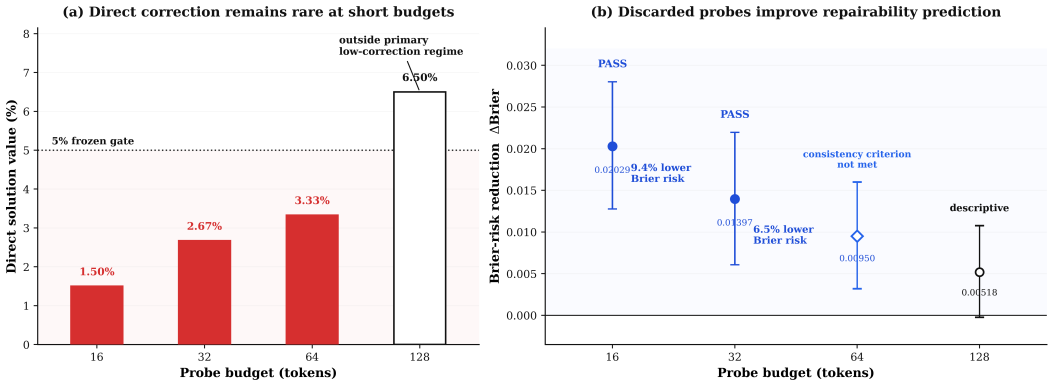

Figure 4: **Primary diagnostic-before-corrective result.** (a) Direct correction remains rare at the three primary short budgets; 128 tokens is descriptive and exceeds the 5% correction gate. (b) Discarded probes improve future-repairability prediction. Budgets 16 and 32 pass the prespecified gate; 64 tokens fails replicate consistency. Error bars are 95% paired state-bootstrap intervals.

specified operational definition of a low-correction regime, not a universal boundary. Confidence intervals use 10,000 paired stratified state-level bootstrap draws; full construction and inference details are in Appendices C–D.

4 Results

4.1 Diagnostic value appears before substantial correction

Figure 4 summarizes the central result. At 16 tokens, only 1.50% of probes directly correct the baseline error, yet out-of-fold repair-trial Brier risk falls from 0.21551 to 0.19523 ($\Delta\mathrm{Br} = 0.02029$; 95% CI $[0.01277, 0.02803]$); 2/3 replicates are positive and the Holm-adjusted reassignment test gives $p = 0.0003$. In other words, the probe is already informative about the value of future computation at a budget where it rarely fixes the current answer itself. The 32-token budget also passes every prespecified gate ($\Delta\mathrm{Br} = 0.01397$; 3/3 positive replicates).

For this fixed finite observer, empirical pooled $\Delta\mathrm{Br}$ declines across tested budgets; we do not interpret this finite-grid pattern as population monotonicity. At 64 tokens the pooled interval remains positive, but only 1/3 replicates is positive, so the consistency screen fails. At 128 tokens direct correction reaches 6.50% and the pooled interval includes zero. Thus the evidence supports a short-budget separation, not monotonic improvement with probe length.

The smallest passing budget in the prospectively tested primary grid is $b^*_{\mathrm{tested}} = 16$; this is a grid-specific result, not a universal threshold.

Table 2: Exact primary budget results. "Pass" refers to the complete prospectively specified gate, not to pooled significance alone.

b	Direct corr.	ΔBr	95% CI	Positive reps	Status
16	1.50%	0.02029	$[0.01277, 0.02803]$	2/3	Pass
32	2.67%	0.01397	$[0.00607, 0.02196]$	3/3	Pass
64	3.33%	0.00950	$[0.00320, 0.01600]$	1/3	Consistency fail
128	6.50%	0.00518	$[-0.00023, 0.01076]$	2/3	Descriptive

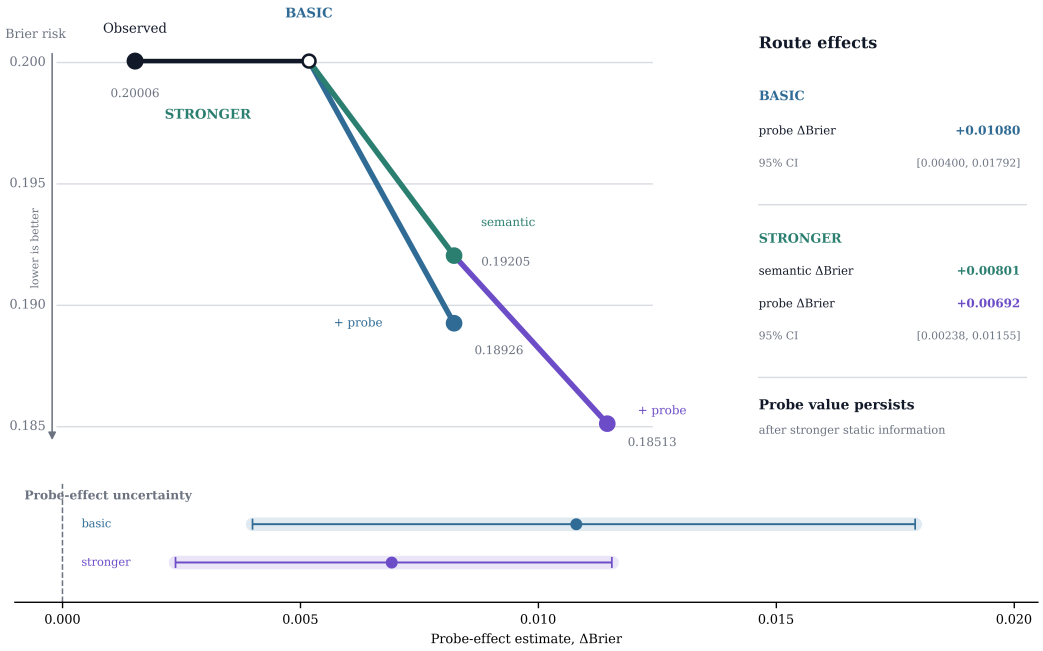

Figure 5: **Robustness of the 16-token diagnostic signal.** The main trajectory compares Brier-risk states under a basic static observer and a stronger frozen semantic observer. The right summary separates the static semantic gain from the incremental discarded-probe gain, and the lower strip reports the paired 95% bootstrap intervals for the probe effect. The probe remains diagnostically valuable after strengthening the pre-probe observer.

4.2 Static difficulty does not exhaust the signal

A natural alternative is that the probe simply recovers information already available from the failed state, especially coarse problem difficulty. Adding MATH category and difficulty improves pre-probe Brier risk from 0.21551 to 0.20006, but the discarded 16-token probe lowers it further to 0.18926 (ΔBr $= 0.01080$, 95% CI $[0.00400, 0.01792]$; 2/3 positive replicates; reassignment $p = 0.0001$). Figure 5 also reports a separately frozen semantic-observer follow-up: the stronger static observer reaches 0.19205 before probing, and the same 16-token probe lowers risk further to 0.18513 (ΔBr $=$ 0.00692, 95% CI $[0.00238, 0.01155]$; 3/3 positive). The follow-up uses frozen state embeddings and nested out-of-fold stacking, so it asks whether the probe adds signal beyond a materially stronger pre-probe representation rather than merely beyond the original sparse baseline; Appendix F gives the full specification and analysis history. Because this P5-B2 protocol was frozen after an earlier semantic-concatenation attempt failed its stronger-baseline requirement, we treat it as a robustness sensitivity rather than part of the original confirmatory gate. These checks reduce, but do not eliminate, static-state explanations.

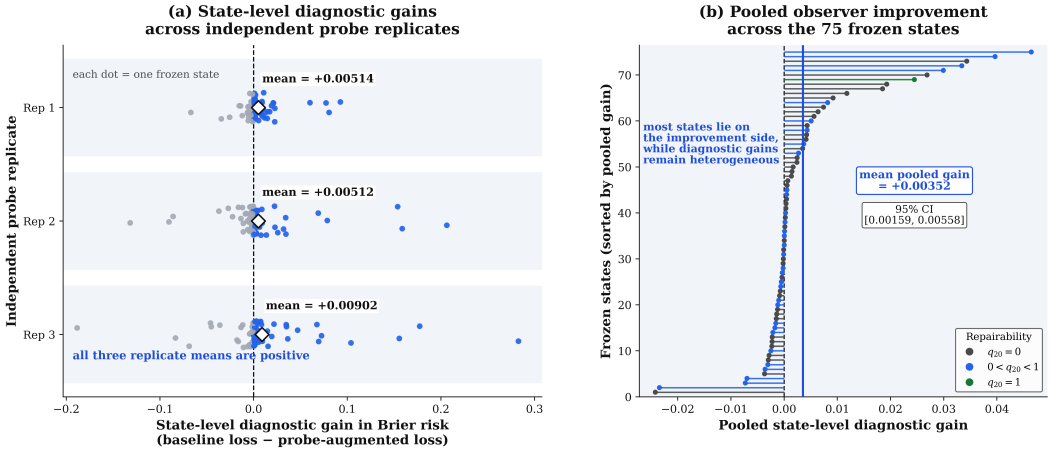

Figure 6: **Cross-family replication on 75 frozen DeepSeek/GSM8K states.** (a) State-level OOF gains for three 16-token replicates (all mean-positive). (b) Pooled $\Delta\mathrm{Br} = 0.00352$ (95% CI $[0.00159, 0.00558]$).

4.3 Cross-family replication preserves the qualitative separation

We repeat the branch-separated test on 75 frozen baseline-wrong GSM8K states (Cobbe et al., 2021) using DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). Twenty independent 512-token repairs per state give mean $\widehat{q}_{20} = 0.1327$ (42 never repaired, 32 intermittent, one always repaired). The historical DeepSeek observer retains its frozen joint-pooling rule across the three probe rows per state, while replicate-specific fits are used only as consistency summaries. Under the historical endpoint operator, none of the 225 16-token probes directly solves the problem. Yet repair-trial Brier risk falls from 0.12422 to 0.12070 ($\Delta\mathrm{Br} = 0.00352$, 95% CI $[0.00159, 0.00558]$), with all three replicate gains positive (0.00514, 0.00512, 0.00902). Figure 6 shows the replicate and pooled gains. The effect is smaller than in Qwen/MATH, so we treat it as cross-family, cross-benchmark support for the qualitative separation, not as evidence of a universal effect size or token threshold.

4.4 Regime dependence: diagnostic value is not a fixed token threshold

Figure 7(a) compares frozen DeepSeek/GSM8K diagnostic results across cohorts and budgets. The 75-state replication is positive at 16 tokens ($\Delta\mathrm{Br} = 0.00352$), whereas the frozen 106-state extension is near zero and slightly negative (-0.00045, 95% CI $[-0.00369, 0.00253]$). Thus the same nominal 16-token budget does not imply the same diagnostic regime, even within one model family and benchmark.

A fresh 153-state cohort evaluated with the exact historical intervention and pooled observer is positive at 8 and 16 tokens ($\Delta\mathrm{Br} = 0.00451$ and 0.00625; 95% CIs $[0.00195, 0.00721]$ and $[0.00191, 0.01071]$), but not at 32 tokens (-0.00044, CI $[-0.00472, 0.00384]$). Post-hoc within-fold probe reassignment supports the 8- and 16-token associations after Holm correction ($p = 0.032, 0.024$), not 32 tokens ($p = 0.989$); we treat this as pairing-robustness evidence rather than a new confirmatory gate pass. A source-code audit found different direct-correction readout operators across the DeepSeek cohorts, so we do not compare their solution-value rates on a common scale; Appendix H records that audit. Taken together, the evidence points to a model-, population-, and budget-dependent diagnostic regime rather than a fixed token threshold. The boundary result is therefore part of the finding, not an exception to it.

4.5 Diagnostic value does not guarantee decision value

Information can be useful without being worth buying. This is the operational difference between diagnostic value and decision value. Figure 7(b) tests the distinction under a compute-matched allocation rule that charges the probe against the same total token budget available to the baseline. At a total budget equivalent to k 512-token repairs, the baseline policy funds k repairs, whereas

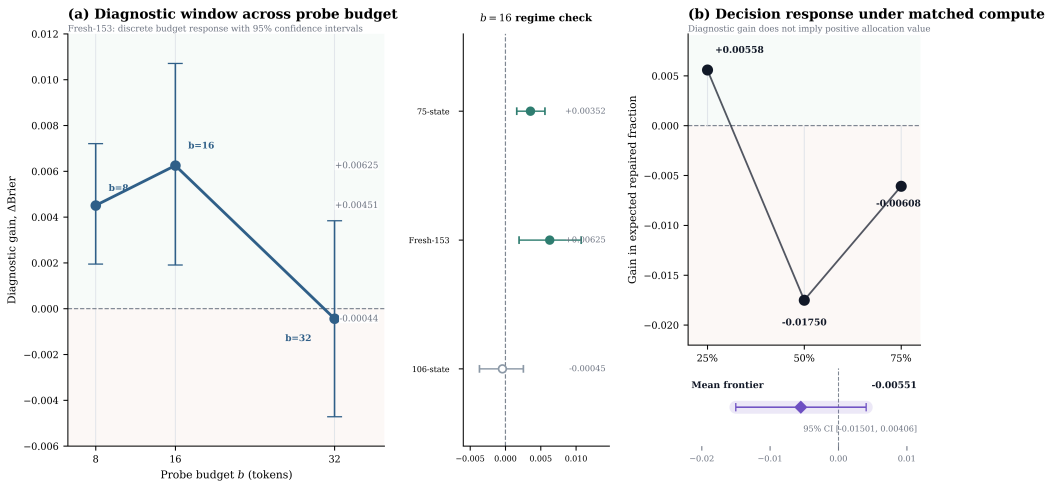

Figure 7: **Boundary and decision-value evidence.** (a) Fresh-153 diagnostic gain across the frozen 8-, 16-, and 32-token budgets with pointwise 95% CIs, together with a same-budget $b = 16$ cross-regime check. Protocol-specific direct-solution rates are deliberately not placed on a common axis. (b) Compute-matched allocation gain at the 25%, 50%, and 75% continuation anchors; the lower certificate reports the mean frontier and its 95% CI.

the diagnostic policy first pays for one 16-token probe per state and therefore funds only $\lfloor (512k - 16N)/512 \rfloor$ repairs. At the 25%, 50%, and 75% anchors ($k = 50, 100, 150$), the diagnostic quotas are $43, 93, 143$ and the gains in expected repaired fraction are $+0.00558$, -0.01750, and -0.00608. Across the full 101-budget frontier, the mean gain is -0.00551 (95% CI $[-0.01501, 0.00406]$). Thus the probe can improve repairability prediction while the tested controller fails to convert that information into uniformly positive policy value. The result does not negate diagnostic value; it shows that the selectivity gain from better information must be large enough to offset the computation spent acquiring it.

5 Related Work and Discussion

Adaptive compute and future value. Adaptive-compute methods allocate inference using input/budget signals, retained prefixes, value forecasts, residual opportunity, solvability, and confidence (Damani et al., 2025; Manvi et al., 2024; 2026; Zabounidis et al., 2025; Wang et al., 2026; Zhou et al., 2026; Yang et al., 2026; Fu et al., 2026). Our contribution is an identification test: can a discarded short reasoning branch measure future compute value? Probe-and-discard isolates this question by restarting repair from the pre-probe state. Appendix M compares protocols.

Correctness and static verification. A second literature asks whether current reasoning is correct or promising. Self-verification, calibration, capability-boundary signals, retained-prefix value models, process reward models, and influence analyses extract signals from current or retained reasoning (Zhang et al., 2025a; Yoon et al., 2025; Zhang et al., 2025b; Chen et al., 2025; Lightman et al., 2023; Wang et al., 2024; Bigelow et al., 2025; Bogdan et al., 2025; Ton et al., 2025; Qian et al., 2025). These approaches are complementary: a verifier can score the frozen trace before any probe. Our target is narrower—information in discarded reasoning about an independently restarted repair, not correctness of the current trajectory. Proposition 4 formalizes the distinction.

Scope. The evidence supports diagnostic computation as a regime, not a universal token rule. The effect depends on observer, model, population, and budget; repairability is intervention-specific; and allocation gains are not uniformly positive. Static controls do not identify a causal semantic mechanism. Thus the experiments establish information value, not automatic deployment value. Appendix L states the corresponding non-claims.

6 CONCLUSION

Test-time computation can serve two roles: advancing an answer and measuring whether more computation is worth spending. Across frozen Qwen/MATH and DeepSeek/GSM8K experiments, short discarded probes predict independently restarted repair success before substantial direct correction, while later cohorts show that this regime is model-, population-, and budget-dependent. Compute-matched allocation further shows that useful information is not automatically worth purchasing.

The broader implication is a *measure–decide–act* view of adaptive inference: solution value, diagnostic value, and decision value are distinct, and practical benefit depends on whether improved selectivity is large enough to justify probe cost.

Reproducibility statement. Confirmatory populations, folds, budgets, observer pipelines, statistical gates, seeds, scoring, and artifacts were frozen; the appendix documents the fresh-cohort protocol correction, post-hoc robustness analysis, and negative-result history, while complete checksum and provenance records are retained in the supplementary artifact manifest.

AI Use Statement. Generative AI tools were used for literature discovery and synthesis; brainstorming and refinement of research questions; feedback on research methodology and experimental design; drafting, revision, and language polishing; assistance with software code; assistance in formulating and checking mathematical exposition, claims, and proof arguments; and support in interpreting experimental outputs. The authors reviewed and verified the experimental protocols, numerical results, citations, mathematical claims and proofs, and final conclusions, and take responsibility for the complete content of the submission.

REFERENCES

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b6a8ca3d06ffa2e044bff8c723863ff-Abstract-Conference.html.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=VnSlfeRCaU.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

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. *arXiv preprint arXiv:2407.21787*, 2024. doi: 10.48550/arXiv.2407.21787. URL https://arxiv.org/abs/2407.21787.

Wenxiang Chen, Wei He, Zhiheng Xi, Honglin Guo, Boyang Hong, Jiazheng Zhang, Nijun Li, Tao Gui, Yun Li, Qi Zhang, and Xuanjing Huang. Better process supervision with bi-directional rewarding signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14471–14485. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.747. URL https://aclanthology.org/2025.findings-acl.747/.

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. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of lm computation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ff414825df833edb8b1839e3d5d495e9-Abstract-Conference.html.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/98257285340854262185500e59bc0f28-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 346–355, 2012. URL https://www.auai.org/uai2012/papers/123.pdf.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733.

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. *arXiv preprint arXiv:2305.20050*, 2023. doi: 10.48550/arXiv.2305.20050. URL https://arxiv.org/abs/2305.20050.

Rohin Manvi, Anikait Singh, and Stefano Ermon. Adaptive inference-time compute: Llms can predict if they can do better, even mid-generation. *arXiv preprint arXiv:2410.02725*, 2024. doi: 10.48550/arXiv.2410.02725. URL https://arxiv.org/abs/2410.02725.

Rohin Manvi, Joey Hong, Tim Seyde, Maxime Labonne, Mathias Lechner, and Sergey Levine. Zero-overhead introspection for adaptive test-time compute. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b22c5b6a87d259769fd186b016412f6a-Abstract-Conference.html.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0421. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/12867c8f1888d97d4e9f95f03c0f3ec9-Abstract-Conference.html.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991. doi: 10.1016/0004-3702(91)90015-C.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59784–59811. PMLR, 2025. URL https://proceedings.mlr.press/v267/ton25a.html.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Ziming Wang, Ivor Tsang, and Hangwei Qian. Aera: Adaptive evidence residual allocation for efficient test-time reasoning. *arXiv preprint arXiv:2608.27964*, 2026. doi: 10.48550/arXiv.2608.27964. URL https://arxiv.org/abs/2608.27964.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8df90a1440ce782d1f5607b7a38f2531-Abstract-Conference.html.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3469. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/95ee7181fe256439dca74ec794b33bb6-Abstract-Conference.html.

Renos Zabounidis, Aditya Golatkar, Michael Kleinman, Alessandro Achille, Wei Xia, and Stefano Soatto. Re-forc: Adaptive reward prediction for efficient chain-of-thought reasoning. In *NeurIPS 2025 Workshop: Efficient Reasoning*, 2025. URL https://openreview.net/forum?id=HUTjOEngpC.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=O6I0Av7683.

Qingjie Zhang, Yujia Fu, Yang Wang, Liu Yan, Tao Wei, Ke Xu, Minlie Huang, and Han Qiu. Stop before you fail: Operational capability boundaries for mitigating unproductive reasoning in large reasoning models. *arXiv preprint arXiv:2509.24711*, 2025b. doi: 10.48550/arXiv.2509.24711. URL https://arxiv.org/abs/2509.24711. Version 4, revised August 2026.

Zhaomeng Zhou, Lan Zhang, Junyang Wang, Mu Yuan, and Junda Lin. Nice fold or hero call: Learning budget-efficient thinking for adaptive reasoning. *arXiv preprint arXiv:2605.11625*, 2026. doi: 10.48550/arXiv.2605.11625. URL https://arxiv.org/abs/2605.11625.

A Reviewer Roadmap and Claim–Evidence Map

The paper is organized around one identification question: whether a short, discarded sample of ordinary reasoning provides observer-relative information about the success of an independent future-repair intervention before that short sample becomes substantially corrective. Table 3 maps each claim to the experiment intended to support it.

Table 3: Claim–evidence map. The paper intentionally separates evidence for solution, diagnostic, and decision value.

Claim	Evidence
Future repairability is heterogeneous	200-state $\widehat{q}_{20}$ distribution; Fig. 3
Short computation can be diagnostically useful before substantial correction	Frozen Qwen/MATH primary; Fig. 4, Table 2
Coarse static difficulty does not exhaust the signal	Category+difficulty robustness; Fig. 5, Appendix F
Follow-up signal persists under a stronger frozen static observer	Separately frozen P5-B2 follow-up after failed P5-B v1; Fig. 5, Appendix F
Qualitative separation replicates across model family/benchmark	DeepSeek/GSM8K 75-state replication; Fig. 6
Diagnostic regime is model-, population-, and budget-dependent	Primary Qwen result (Table 2; Fig. 4) plus 75-state replication, 106-state extension, and fresh 153-state cohort; Fig. 7(a)
Diagnostic value does not guarantee realized policy value	Compute-matched allocation analysis; Fig. 7(b)

B Definitions, Identities, and Supporting Results

Proposition 1 (Observer-relative diagnostic-value characterization)**.** *Let $r(S) = \Pr(Y = 1 \mid S)$ denote the repair-success probability associated with the frozen generating state, and let $O = \phi(S)$. If the probe and evaluated repair are conditionally independent given the frozen state,*

$$Z_b \perp Y \mid S,$$

then

$$\eta_0(O) = \mathbb{E}[r(S) \mid O], \qquad \eta_b(O, Z_b) = \mathbb{E}[r(S) \mid O, Z_b].$$

Consequently,

$$D_b^{\mathrm{Br}}(O) = \mathbb{E}[(\eta_b - \eta_0)^2] \tag{8}$$

$$= \mathbb{E}[\mathrm{Var}(\eta_b(O, Z_b) \mid O)] \geq 0. \tag{9}$$

Moreover, $D_b^{\mathrm{Br}}(O) > 0$ if and only if $\Pr\{\eta_b(O, Z_b) \neq \eta_0(O)\} > 0$. Thus branch separation is fully compatible with positive diagnostic value: the probe need not cause repair success; it need only reveal repair-relevant variation in S that is not already resolved by O.

Proof. Because O is a function of S and $Z_b \perp Y \mid S$,

$$\mathbb{E}[Y \mid O, Z_b] = \mathbb{E}[\mathbb{E}[Y \mid S, O, Z_b] \mid O, Z_b] = \mathbb{E}[r(S) \mid O, Z_b],$$

which gives the expression for η_b. The same tower-property argument gives $\eta_0(O) = \mathbb{E}[r(S) \mid O]$. Since $\eta_0 = \mathbb{E}[\eta_b \mid O]$, conditional-expectation orthogonality yields

$$\mathbb{E}[(Y - \eta_0)^2] = \mathbb{E}[(Y - \eta_b)^2] + \mathbb{E}[(\eta_b - \eta_0)^2].$$

Finally,

$$\mathbb{E}[(\eta_b - \eta_0)^2] = \mathbb{E}\big[\mathbb{E}\big[(\eta_b - \mathbb{E}[\eta_b \mid O])^2 \mid O\big]\big] = \mathbb{E}[\mathrm{Var}(\eta_b \mid O)].$$

The last quantity is zero exactly when $\eta_b = \eta_0$ almost surely, and is strictly positive otherwise. □

Proposition 2 (Finite-K repairability-risk decomposition)**.** *Let $Y_1, \ldots, Y_K$ be conditionally independent Bernoulli repair outcomes with $\Pr(Y_k = 1 \mid S) = r(S)$, and define $\widehat{q}_K = K^{-1} \sum_{k=1}^{K} Y_k$. Let p be any prediction constructed from observer information whose randomness is independent of the repair-rollout randomness conditional on S. Then*

$$\mathbb{E}\big[(\widehat{q}_K - p)^2 \mid S, p\big] = (r(S) - p)^2 + \frac{r(S)(1 - r(S))}{K}. \tag{10}$$

Hence, for any two such predictions p_0 and p_1,

$$\begin{aligned} &\mathbb{E}\big[(\widehat{q}_K - p_0)^2 - (\widehat{q}_K - p_1)^2\big] \\ &\quad = \mathbb{E}\big[(r(S) - p_0)^2 - (r(S) - p_1)^2\big]. \end{aligned} \tag{11}$$

Moreover, for the realized K repair trials,

$$\frac{1}{K} \sum_{k=1}^{K} (Y_k - p)^2 = (\widehat{q}_K - p)^2 + \widehat{q}_K(1 - \widehat{q}_K). \tag{12}$$

Therefore the repair-trial Brier-risk difference between two observers is exactly the same, sample by sample, as their soft-label squared-error difference.

Proof. Conditional on (S, p), the assumed conditional independence of the prediction from the repair-rollout randomness gives $\mathbb{E}[\widehat{q}_K \mid S, p] = r(S)$ and $\mathrm{Var}(\widehat{q}_K \mid S, p) = r(S)(1 - r(S))/K$. The bias–variance decomposition gives Eq. (10). Subtracting the expressions for p_0 and p_1 cancels the finite-K variance term; applying the law of total expectation then gives Eq. (11). For Eq. (12), expand

$$\frac{1}{K} \sum_{k=1}^{K} (Y_k - p)^2 = \frac{1}{K} \sum_{k=1}^{K} Y_k - 2p\widehat{q}_K + p^2 = \widehat{q}_K - 2p\widehat{q}_K + p^2,$$

because $Y_k^2 = Y_k$. Expanding $(\widehat{q}_K - p)^2 + \widehat{q}_K(1 - \widehat{q}_K)$ gives the same expression. □

Use in the empirical analysis. Equations (10)–(11) require the stated conditional-independence condition on the prediction. They are not used to justify the empirical confidence intervals. The realized-sample identity in Eq. (12) is purely algebraic and holds for any fixed prediction, so the reported observer-to-observer $\Delta\mathrm{Br}$ is exactly the same under repair-trial Brier loss and soft-label squared error. Sampling uncertainty is handled separately by the paired state-level bootstrap conditional on the frozen out-of-fold prediction pipeline described in Appendix D.

Proposition 3 (Gross and net decision value under exact posteriors)**.** *Let $\eta_0 = \Pr(Y = 1 \mid O)$ and $\eta_b = \Pr(Y = 1 \mid O, Z_b)$ be the exact conditional success probabilities under the nested information sets, and suppose a successful long continuation has value $v > 0$ and cost $c_L \geq 0$. Define $h(u) = (vu - c_L)_+$. Then*

$$\mathbb{E}[h(\eta_b)] - \mathbb{E}[h(\eta_0)] \geq 0. \tag{13}$$

Moreover, with decision threshold $t = c_L/v$, the inequality is strict if and only if, for a set of O-values of positive probability, the conditional distribution of η_b given O places positive probability on both sides of t. Thus positive diagnostic value need not imply strictly positive gross decision value: posterior refinement can improve prediction without changing any action. If obtaining the probe costs $c_b \geq 0$, the net value

$$\mathrm{NVDC}_b = \mathbb{E}[h(\eta_b)] - \mathbb{E}[h(\eta_0)] - c_b$$

can be negative. No analogous nonnegativity guarantee holds for a learned finite-sample predictor combined with a fixed heuristic allocation rule.

Proof. By the tower property, $\eta_0 = \mathbb{E}[\eta_b \mid O]$. The function $h(u) = (vu - c_L)_+$ is convex and is affine on each side of $t = c_L/v$, so conditional Jensen's inequality gives

$$h(\eta_0) = h(\mathbb{E}[\eta_b \mid O]) \leq \mathbb{E}[h(\eta_b) \mid O].$$

Taking expectations proves Eq. (13). Because h is affine on each side of the kink at t, the conditional Jensen gap is zero when $\eta_b \mid O$ lies almost surely in one affine region and is strictly positive exactly when its conditional distribution places positive probability on both sides of t for a set of O-values of positive probability. Subtracting the nonnegative acquisition cost c_b need not preserve nonnegativity. The final statement follows because estimated probabilities and finite-sample ranking rules need not satisfy the nested conditional-expectation identities used above. □

Connection to fixed-budget batch allocation. Proposition 3 is a pointwise value-of-information result. For a *fixed* repair quota m, ranking by exact richer posteriors can only weakly improve the Bayes-optimal expected yield because the top-m set maximizes the posterior sum under the available information. In the compute-matched experiment, however, buying probes reduces the feasible repair quota from m_0 to $m_b < m_0$. No dominance follows across different quotas: the selectivity gain from better ranking can be smaller than the structural capacity loss from purchasing probes. This is why positive diagnostic value and even positive gross pointwise value do not imply a positive compute-matched frontier.

Proposition 4 (Correctness information does not identify repair information)**.** *Let A denote a correctness event distinct from the fixed future-repair outcome Y—for example, terminal correctness of the probe branch. (The empirical population's baseline correctness is already fixed at zero by construction.) Without additional structural assumptions, neither*

$$I(A; Z_b \mid O) > 0 \implies I(Y; Z_b \mid O) > 0$$

nor

$$I(Y; Z_b \mid O) > 0 \implies I(A; Z_b \mid O) > 0$$

holds in general, even while preserving the paper's branch-separation condition $Z_b \perp Y \mid S$.

Proof. Let the frozen latent state be $S = (U, V)$, where U and V are independent fair Bernoulli variables, and let O be constant. First set $Z_b = U$, $A = U$, and $Y = V$. Then $I(A; Z_b \mid O) = 1$ bit while $I(Y; Z_b \mid O) = 0$ bits. Conversely, set $Z_b = U$, $A = V$, and $Y = U$. Then $I(A; Z_b \mid O) = 0$ bits while $I(Y; Z_b \mid O) = 1$ bit. In both constructions, Z_b and Y are deterministic once S is given, so their conditional joint law factorizes and $Z_b \perp Y \mid S$ holds. Thus branch separation alone does not identify correctness information with future-repair information in either direction. □

Interpretation for the empirical estimand. Propositions 1–4 separate four issues that are easy to conflate. Branch separation removes token carry-over without requiring the probe to be statistically independent of the latent state. Finite-K repair sampling adds a prediction-independent variance term that cancels from observer-to-observer squared-risk differences; it does not remove state-sampling uncertainty, which is why inference still resamples whole states with all attached repairs. Positive diagnostic information is a statement about prediction, whereas positive deployment value additionally depends on probe cost, calibration, and the downstream controller. Finally, information about answer correctness and information about future repair are logically distinct without additional structural assumptions. These distinctions motivate the paper's separation

$$\text{solution value} \neq \text{diagnostic value} \neq \text{decision value}.$$

C Complete Primary Protocol

Population. The confirmatory primary population contains 200 Qwen3-8B/MATH states whose frozen baseline attempt is incorrect. Population construction is state-level: all repair rollouts and all probe replicates from one state remain tied to that state in folds, bootstrap resampling, and diagnostic evaluation.

Frozen branch intervention and decoding. The primary branches use Qwen3-8B revision `b968826d...1218` and one locked revision template (SHA256 `a0fe8afa...d05d817`) containing exactly one problem field, frozen reasoning trace, and current-answer field. The system instruction asks for an additional verification/revision pass that checks the existing reasoning, repairs errors when possible, and preserves an already-correct answer; the user block supplies `PROBLEM`, `PREVIOUS REASONING TRACE`, and `CURRENT ANSWER`, followed by "Perform one additional reasoning/revision pass." Both repairs and probes use stochastic decoding with temperature 0.7, top-$p = 0.9$, and EOS or the prespecified token cap as the stopping rule. Repairs use a 512-token cap and seed namespace 610000; probes use caps $b \in \{16, 32, 64, 128\}$ and a disjoint 710000 seed namespace. Hence the two branch types differ in horizon and independent randomness, not in the frozen state or intervention semantics.

Future repair target. Each state receives $K = 20$ independently sampled continuations with a 512-token maximum. Let $Y_{ik} \in \{0, 1\}$ denote correctness of repair k for state i and $\widehat{q}_{20,i} = 20^{-1} \sum_k Y_{ik}$. The realized summary is: 200 states; 1,258 successful repairs among 4,000 continuations; mean $\widehat{q}_{20} = 0.3145$; 77 states at zero; 101 strictly between zero and one; and 22 at one.

Probe generation. For each state and each $b \in \{16, 32, 64, 128\}$, three independent probes are generated from the same pre-probe state. The probe branch and each long-repair branch have independent sampling randomness. Probe text is observed only by the augmented diagnostic observer and is not appended to the evaluated repair branch.

Direct solution scoring. For the primary Qwen experiment, direct solution value is an endpoint property of the newly emitted probe branch. The frozen scorer applies a deterministic 256-token answer readout (sampling disabled; thinking disabled) to the probe output and then the deterministic correctness rule. Readout tokens are scoring infrastructure: they are not appended to the probe, are not visible to the diagnostic observer, and are not counted in the probe budget b. No diagnostic-observer output is used to determine whether a probe solves the problem. Consequently, a short probe can contain useful intermediate progress without being counted as a direct correction unless it supports the correct terminal answer at this readout. This is a protocol-specific endpoint measure of probe-supported correctness, not a claim that the correct answer literally appears within the first b emitted tokens.

Primary family. Budgets 16, 32, and 64 form the confirmatory multiplicity family. Budget 128 is descriptive. The low-correction gate is 5%. The complete primary gate requires positive pooled Brier reduction, Holm-adjusted significance in the full-pipeline reassignment test, an operational replicate-consistency screen requiring at least two of three positive replicate-specific gains, and direct solution value no greater than 5%. The 2-of-3 requirement is a consistency screen rather than an additional significance test, and the 5% threshold is a prespecified operational definition of a low-correction regime rather than a universal boundary.

D Observer Specification and Statistical Inference

Frozen folds. Five state-level folds are fixed before probe-outcome evaluation. All probe replicates from a state inherit the state's fold. Predictions reported in the paper are out of fold.

Primary representation and weighted binomial fitting. The baseline text consists of problem statement, frozen failed reasoning, and frozen failed answer. The augmented text adds the diagnostic probe. The frozen feature stack uses word 1–2 grams and character 3–5 grams with TF–IDF normalization, followed by ℓ_2-regularized logistic regression with $C = 1$ and the `liblinear` solver. No hyperparameter search is performed on diagnostic outcomes. If state i has s_i successful repairs among $K = 20$, training uses two copies of its feature vector with binary labels 1 and 0 and sample weights s_i and $K - s_i$, respectively. Thus `liblinear` is fit to a weighted binary likelihood equivalent to expanding the 20 repair outcomes; $\widehat{q}_{20,i}$ is not passed as a continuous class label. For the primary Qwen pooled augmented fit, the three probe rows per state are stacked jointly, the two binomial weights attached to each repeated row are divided by three so the total target weight remains K, and the three held-out probabilities are averaged to one state-level prediction. Separate replicate-specific fits supply the consistency screen. The historical DeepSeek analyses retain their originally frozen joint-pooling specification, with replicate-specific fits reported separately as consistency summaries.

Repair-trial Brier risk and soft-label equivalence. The diagnostic target is the empirical repair probability $\widehat{q}_{20,i}$. Using a soft label is equivalent to averaging binary Brier losses over the 20 repair outcomes up to a term independent of the prediction: for any prediction p_i,

$$\frac{1}{20} \sum_{k=1}^{20} (Y_{ik} - p_i)^2 = (\widehat{q}_{20,i} - p_i)^2 + \widehat{q}_{20,i}(1 - \widehat{q}_{20,i}). \tag{14}$$

The absolute Brier levels reported in the experiments use the repair-trial (Bernoulli-expanded) form on the left-hand side. The second term is independent of the prediction, so the reduction $\Delta\mathrm{Br}$ between two observers is exactly equal to the corresponding reduction in soft-label squared error.

Bootstrap. Confidence intervals use 10,000 paired state-level bootstrap draws stratified by the frozen $\widehat{q}_{20}$ stratum used in fold construction. States are resampled within stratum, and each sampled state carries its baseline and augmented predictions together, preserving the paired comparison and the empirical repairability mix. Observer models are not refit inside bootstrap draws; these intervals therefore quantify state-resampling uncertainty conditional on the frozen out-of-fold prediction pipeline, not full-pipeline retraining uncertainty.

Full-pipeline reassignment test. The permutation-style test reassigns probe instances within the frozen fold-compatible structure and reruns the prespecified augmented fitting/evaluation pipeline. This targets whether the observed association depends on the correct state–probe pairing rather than on generic additional text. We interpret this as a state–probe pairing robustness test; it is not an exact test of $Z_b \perp Y \mid O$, because reassignment disrupts the probe's relationship to the underlying state as well as to the repair outcome. The primary family is adjusted by Holm's procedure (Holm, 1979).

E Complete Primary Budget Results

Table 4: Complete frozen primary budget audit. Budget 128 was prespecified descriptive and excluded from the Holm-adjusted primary family.

b	Direct	ΔBr	95% CI	Positive reps	Holm p	Gate
16	0.0150	0.02029	$[0.01277, 0.02803]$	2/3	0.0003	Pass
32	0.0267	0.01397	$[0.00607, 0.02196]$	3/3	0.0003	Pass
64	0.0333	0.00950	$[0.00320, 0.01600]$	1/3	0.0003	Fail consistency screen
128	0.0650	0.00518	$[-0.00023, 0.01076]$	2/3	–	Descriptive

The exact replicate-specific gains at the primary budgets are retained in the frozen result artifacts. The gate is conjunctive: a positive pooled estimate or small pooled p-value cannot rescue a budget that fails replicate consistency or the direct-correction requirement.

Length-only negative control. A prespecified length-only control yields essentially zero incremental gain (approximately -9.8×10^{-5} in Brier reduction). Since nominal budget is fixed within each budget-specific comparison, the primary diagnostic improvement cannot be attributed simply to observing a varying nominal token count.

F Robustness: Static Controls and Stronger Observer

Category and difficulty controls. The frozen static Brier risks are 0.21551 for the primary baseline, 0.20566 after adding category, 0.21202 after adding difficulty, and 0.20006 after adding both. With category+difficulty included, the 16-token augmented risk is 0.18926, giving $\Delta\mathrm{Br} = 0.0107991$. The 95% CI is $[0.003996, 0.017921]$, replicate-specific gains are $+0.00657, +0.01093, -0.00743$, and the 10,000-draw full-pipeline reassignment test gives one-sided $p = 0.0001$. Direct solution value is unchanged at 0.015.

Separately frozen stronger-observer follow-up. The reported stronger-observer result is the P5-B2 follow-up, not the preceding P5-B feature-concatenation attempt. P5-B first froze 768-dimensional state embeddings from `sentence-transformers/all-mpnet-base-v2` using a maximum input length of 384 tokens, attention-mask mean pooling of the last hidden layer, and ℓ_2 normalization. That initial concatenation made the static observer worse and was stopped as a failed robustness design; after observing that failure, the P5-B2 protocol was written and frozen before its outcomes were inspected.

P5-B2 reuses those frozen state embeddings but keeps the original P5-A category+difficulty predictor intact. Within each of the five outer folds, the 768-dimensional embedding is reduced to four PCA components and standardized using training data only. Deterministic four-fold inner out-of-fold predictions from the P5-A branch and the semantic branch are convexly stacked with weights on a 0.05 grid, chosen by inner Brier risk with ties favoring the lower semantic weight. The selected

semantic weights across outer folds are 0.35, 0.40, 0.50, 0.30, and 0.35. This safe stack reduces static Brier risk from 0.200058 to 0.192045. Only after this strength gate passes is the 16-token probe evaluated: inner out-of-fold stacking then combines the P5-A static, semantic, and pooled probe predictors on the corresponding 0.05 simplex grid. The final augmented risk is 0.185125, giving incremental $\Delta\mathrm{Br} = 0.006920$ with 95% CI $[0.002376, 0.011547]$ and 3/3 positive replicate-specific gains. Because P5-B2 was designed after the failed P5-B attempt, it is reported as a separately frozen follow-up sensitivity and is not substituted for the original confirmatory estimand.

G Cross-Family Replication

The DeepSeek-R1-Distill-Llama-8B/GSM8K replication freezes 75 baseline-wrong states. Twenty independent 512-token repairs per state yield 1,500 repair attempts and mean $\widehat{q}_{20} = 0.1327$: 42 states have zero successful repairs, 32 are intermediate, and one is repaired in all 20 trials. Three 16-token probes per state yield 225 probe runs, with no observed direct solutions under the historical endpoint operator. That operator reads only the newly emitted probe text, closes DeepSeek's thinking block immediately, uses deterministic decoding capped at 256 tokens, and permits no fallback continuation. Under the frozen historical specification, the pooled observer is fit jointly on the three probe rows per state; replicate-specific fits are reported separately as consistency summaries rather than averaged to form the pooled estimate. The baseline and augmented Brier risks are 0.12421858 and 0.12069841, respectively, so pooled $\Delta\mathrm{Br} = 0.00352017$. The three replicate-specific gains are 0.00514213, 0.00511654, and 0.00901676. A 10,000-draw paired state bootstrap gives 95% CI $[0.001585, 0.005585]$.

The benchmark had appeared in earlier feasibility work, so we do not describe the benchmark itself as untouched. What is prospectively locked for this replication is the 75-state wrong population and the replication protocol before its replication diagnostic outcome.

H Prospectively Frozen Boundary Extension

A later DeepSeek/GSM8K extension prospectively freezes a distinct 106-state baseline-wrong population at the same nominal 16-token probe budget. It yields exactly 2,120 planned long-repair attempts (20 per state). The frozen historical parser resolves 1,411 and leaves 709 unresolved: 707 hit the 512-token cap without an admissible terminal readout, one completed without a numeric answer, and one completed with an explicit nonnumeric answer. Crucially, these 709 attempts are *not excluded*. The frozen target is successful planned repairs divided by 20, so every unresolved attempt remains in the fixed denominator as a non-success. There are 236 successful repairs, mean $\widehat{q}_{20} = 0.111321$, 59 states with zero repair success, 47 intermediate states, and none with $\widehat{q}_{20} = 1$.

The frozen baseline and augmented Brier risks are 0.10106063 and 0.10151042, respectively, so $\Delta\mathrm{Br} = -0.00044979$. The replicate-specific gains are all negative (-0.00102, -0.00789, -0.00322), and the 95% bootstrap interval is $[-0.00369, 0.00253]$.

Endpoint-readout convention. The later 106-state direct-solution audit supplied the problem together with the frozen baseline trace plus newly emitted probe text to a deterministic 256-token readout, with a 1024-token fallback on primary cap hits, and obtained 39/318 correct readouts (0.122642). The historical 75-state operator is materially narrower: it supplies only the newly emitted probe text, explicitly closes DeepSeek's thinking block before decoding, uses a deterministic 256-token readout, and has no fallback. Because these are different solution-value measurement operators, the raw rates are protocol-specific and are not compared numerically. We do not replace either frozen outcome with a post-hoc harmonized solution-value estimate; instead, the 106-state cohort is used for its frozen diagnostic result and as a documented boundary case. This preserves the original protocol history while avoiding a spurious 0% versus 12.26% comparison.

This cohort is retained as a boundary result rather than post-hoc optimized. In particular, we do not search a shorter budget on the same 106 states and relabel it confirmatory. The result directly limits the interpretation of $b^*_{\mathrm{tested}} = 16$: it is a primary-grid statement, not a universal token threshold. The larger fresh-cohort audit below further supports a regime-dependent rather than fixed-token interpretation.

I Fresh 153-State Historical-Observer Audit

Protocol alignment. An earlier fresh-cohort run used a prompt that omitted the current-answer field present in the historical intervention. We therefore regenerated both repair and probe branches under the exact historical prompt before conducting the analyses reported here. Because outcomes from the earlier run had already been inspected, we treat the aligned analysis as a protocol-correction robustness audit rather than untouched prospective confirmation. The regenerated aligned branches were then frozen; the observer sensitivity and reassignment analyses below use those frozen aligned outputs and require no further LLM generation.

Fresh cohort and historical observer. The aligned audit evaluates 153 baseline-wrong DeepSeek states on GSM8K using the exact historical intervention and historical pooled-observer specification. Each state receives 20 independent repairs, giving mean empirical repairability $\widehat{q}_{20} = 0.20719$: 72 states are never repaired, 77 are repaired intermittently, and 4 are repaired in all 20 trials. The later full-branch endpoint readout gives protocol-specific correction rates of 11.33%, 13.94%, and 14.38% at budgets 8, 16, and 32, respectively; these values are retained as outputs of that later readout protocol but are not used for cross-cohort solution-value claims or a low-correction gate. The historical pooled-observer sensitivity uses the frozen probe texts, $\widehat{q}_{20}$ targets, and folds without new LLM generation. Its baseline repair-trial Brier risk is 0.177624. Diagnostic gains are $\Delta\mathrm{Br} = 0.00451$ at 8 tokens, 0.00625 at 16 tokens, and -0.00044 at 32 tokens, with 95% bootstrap CIs $[0.00195, 0.00721]$, $[0.00191, 0.01071]$, and $[-0.00472, 0.00384]$.

Post-hoc pairing robustness. The three probes attached to each state are reassigned as a bundle within the frozen folds, and the exact historical pooled observer is refit. One-sided Monte Carlo p-values from 1,000 reassignments are Holm-adjusted across the three budgets, yielding adjusted $p = 0.03197$, 0.02398, and 0.98901 at 8, 16, and 32 tokens. We interpret this as a state–probe pairing robustness analysis, not as an exact test of conditional independence given the baseline observer. The positive pooled associations at 8 and 16 tokens are therefore treated as robustness evidence, not new confirmatory gate passes: their replicate-consistency counts are 0/3 and 1/3, respectively, and the direct-correction operator is not comparable to the 75-state endpoint operator.

Artifact integrity. All frozen inputs, folds, probe files, historical out-of-fold predictions, and post-hoc permutation outputs used by this audit are versioned and checksum-verified. Complete SHA256 fingerprints and provenance records are retained in the supplementary artifact manifest rather than reproduced in the paper.

J Decision-Value Analysis

The frozen compute-matched allocation analysis evaluates whether improved repairability prediction translates into realized allocation gain under an explicit shared token budget. Let $N = 200$, long-repair cost $L = 512$, probe cost $b = 16$, and let $k \in \{50, \ldots, 150\}$ denote the baseline repair quota. The common budget is $B_k = kL$. Baseline-only ranks states by the frozen pre-probe score and selects $n_0(k) = k$ states. A diagnostic policy samples exactly one 16-token probe for every state, pays the cohort-wide overhead $Nb = 3{,}200$, and therefore selects

$$n_b(k) = \left\lfloor \frac{B_k - Nb}{L} \right\rfloor$$

long repairs. Thus the three stored probe replicates are three alternative realizations of the same one-probe policy, not three probes charged simultaneously to a state. For each realization, states are ranked by its frozen out-of-fold augmented score; ties are broken deterministically by state index.

For a selected set A, policy performance is the cohort-normalized expected repair yield $G(A) = N^{-1} \sum_{i \in A} \widehat{q}_{20,i}$. Allocation gain is $G(A_b) - G(A_0)$, so its unit is an expected repaired fraction of the 200-state cohort. The reported diagnostic curve averages this gain over the three probe realizations. The 25%, 50%, and 75% anchors mean $k/N = 0.25, 0.50, 0.75$, corresponding to baseline quotas $50, 100, 150$ and diagnostic quotas $43, 93, 143$. Their gains are $+0.00558$, -0.01750, and -0.00608. Averaging over all 101 budgets $k = 50, \ldots, 150$ gives the mean frontier gain -0.00551; a 10,000-draw paired bootstrap resampling states within the frozen repairability strata

gives 95% CI $[-0.01501, 0.00406]$. Replicate frontier means are $-0.00586, -0.00197, -0.00869$. Under the stronger observer the mean is -0.00795. This accounting matches token caps, not wall-clock latency, FLOPs, KV-cache traffic, or deterministic endpoint-readout cost.

In the fresh 153-state cohort, an offline compute-matched ranking analysis provides a complementary operating point. At a 20% repair budget, the 16-token probe-guided ranking yields 7.15 expected repaired states versus 5.35 under baseline-only ranking, a descriptive difference of $+1.80$; the 95% bootstrap CI for the difference is $[-2.65, 3.25]$. This offline analysis does not establish positive deployment value.

These estimates are not used to reinterpret the positive diagnostic result as a failure. They evaluate a different quantity: realized policy value under particular finite-sample allocation rules. The results support the paper's separation between information value and decision value.

K Protocol-Development Firewall and Artifact Integrity

The first MATH population run was found to overlap with states used during answer-infrastructure development. It was designated infrastructure/dry-run evidence only before any future-repair or diagnostic-probe outcomes were generated for the valid primary. The confirmatory primary population was drawn from the untouched tail of the original frozen permutation under a written amendment frozen before those outcomes. No answer extraction, scoring, text representation, classifier, fold, primary budget, or significance rule was modified after observing the valid primary diagnostic outcomes.

Publication-grade stages retain immutable parent artifacts, model revision identifiers, seeds, raw generations, deterministic readout/scoring outputs, state-fold assignments, manifests, and checksum verification. The frozen 200-state primary repair summary used for Figure 3 is reproduced by the audit as 200 states, 1,258 successful repairs, mean $\widehat{q}_{20} = 0.3145$, 77 zero states, 101 intermediate states, and 22 one states. Complete fingerprints and provenance records are retained in the supplementary artifact manifest.

L Scope, Identifiability, and Non-Claims

This appendix clarifies the scope of the empirical claims and the role of the probe-and-discard construction. These qualifications are part of the intended interpretation of the results rather than post hoc exceptions.

Why can a branch-independent probe still be informative? Branch independence prevents the diagnostic probe from *causing* success on the evaluated repair branch. It does not imply statistical independence between the information revealed by the probe and the repair outcome, because both branches originate from the same frozen pre-probe reasoning state. Consequently, a disposable probe may reveal information about latent properties of that state even though its tokens are never carried into the evaluated repair trajectory. Proposition 1 formalizes this distinction.

Is 16 tokens a universal diagnostic budget? No universal token threshold is claimed. The quantity $b^*_{\text{tested}} = 16$ denotes only the smallest budget that satisfies the predeclared gate within the prospectively tested primary budget grid. The separately frozen 106-state DeepSeek extension shows that the same nominal 16-token budget can fall in a qualitatively different regime, while the aligned fresh 153-state audit recovers positive diagnostic association at 8 and 16 tokens but not 32 tokens under the historical observer. We therefore interpret diagnostic computation as a model-, population-, and budget-dependent regime rather than a universal token constant.

Could the effect be explained by problem difficulty alone? Static problem characteristics are predictive of future repairability, but they do not account for the full observed effect. Adding MATH category and difficulty to the frozen observer reduces, but does not eliminate, the incremental predictive value of the 16-token discarded probe. A separately frozen stronger pre-probe observer yields the same qualitative conclusion. These analyses support an interpretation in which the probe contributes information beyond the static controls actually tested; they do not establish that all pre-probe information has been exhausted. In particular, a specialized process verifier or model-internal

uncertainty monitor applied directly to the frozen trace could be complementary or could absorb part of the observed incremental signal.

Could the result be an artifact of probe length? The primary comparisons are conducted at fixed probe budgets, so within each budget the diagnostic observer does not learn from a varying nominal token count. In addition, a prespecified length-only negative control produces essentially no predictive gain. The length-only result rules out nominal token count alone as an explanation for the primary gain; it does not identify which textual, semantic, or trajectory features of the generated probe carry the signal.

Does diagnostic value imply that the probe is worth purchasing? No. Diagnostic value is a predictive quantity, not a deployment guarantee. Our compute-matched allocation analyses do not establish uniformly positive decision value; the fresh-cohort routing point is descriptively positive but its uncertainty interval spans zero. Thus the paper deliberately distinguishes *solution value*, *diagnostic value*, and *decision value*. Whether a diagnostic computation is worth performing in practice additionally depends on the downstream policy, the cost of the probe, and the utility assigned to subsequent computation.

Is the claim one of mutual information with respect to the full model state? No. Diagnostic value is defined relative to a fixed, bounded observer O. Under the probe-and-discard construction, an observer with an omniscient description of the complete generating state would have no need for the probe to recover information already contained in that state. The empirical claim is therefore observer-relative: a short discarded computation improves prediction of future repairability for the frozen observer used in the experiment. We do not identify a model-intrinsic mutual-information quantity for the complete hidden state.

How broadly should the empirical phenomenon be expected to generalize? The evidence supports a qualitative phenomenon, not universal generalization. The primary Qwen3-8B/MATH experiment establishes the prospectively specified effect in one model–benchmark setting, and the 75-state DeepSeek-R1-Distill-Llama-8B/GSM8K experiment provides a cross-family, cross-benchmark replication. The separately frozen 106-state cohort shows that the same short budget need not exhibit the same separation in every population, while the aligned fresh 153-state audit provides additional positive diagnostic association at shorter budgets under the historical observer. We report these results together because regime dependence is part of the empirical finding.

M Detailed Positioning Against Closest Work

Table 5: Closest work and the experimental distinction. "Retained?" is restricted to Yes/No/N/A and asks whether the observed reasoning tokens/state belong to the trajectory whose value or correctness is assessed.

Work	Signal / target	Retained?	Distinction from probe-and-discard
Damani et al.	Reward distribution by input and budget	N/A	Predicts compute benefit from the input; no disposable reasoning measurement
Manvi et al. (2024)	Explicit "can I do better?" evaluation	No	Dedicated self-evaluation can control a restart without becoming part of the restarted trajectory
ZIP-RC / Re-FORC	Future reward and/or cost from current prefix	Yes	Forecasts value from a retained evolving state
DEER / Deep Think with Confidence	Trial confidence / local confidence	Yes	Controls an ongoing trace using accumulated evidence
AERA	Residual opportunity at checkpoints	Yes	Sequential controller; current checkpoint is retained
Lightman et al. / Math-Shepherd	Step-level process verification or reward	Yes	Scores retained candidate reasoning; a static score on the frozen failed trace is a complementary observer
Zhang et al. (self-verification)	Hidden-state answer-correctness probe	Yes	Predicts current/upcoming answer correctness from the retained state, not independent repair success
Yoon et al.	Verbalized confidence through CoT	Yes	Studies calibration of current answer confidence
Zhang et al. (Stop Before You Fail)	Failure/capability-boundary signal	Yes	Monitors an ongoing reasoning trace to predict eventual solved vs. unsolved status
BiRM	Forward value of partial solution	Yes	Learned success probability of retained partial trajectory
Ton et al.	Step information gain about gold answer	Yes	Gold-answer information, not independent-repair information
Qian et al.	MI peaks with gold answer	Yes	Information about the gold answer inside ongoing representations
Forking Paths / Thought Anchors	Counterfactual trajectory influence	Yes	Intervenes on retained trajectories to locate influential tokens/sentences
Russell & Wefald / Hay et al.	Value of computation / metalevel evidence	N/A	Foundational decision theory; no LLM branch-separated solution-vs-diagnostic test
Ours	**Short ordinary reasoning; future repair success**	**No**	**Probe discarded; independent repair restarts from the pre-probe state; direct solution value measured separately**