# How Faithful Is a Truncated Reasoning Trace?
# Measuring and Modeling Biased Feedback in Test-Time Compute Allocation

**Anonymous authors**
Paper under double-blind review

## Abstract

Test-time compute allocation methods must decide how many reasoning tokens to spend on each query. A tempting shortcut is to truncate one long reasoning trace and score each prefix, which yields feedback for every smaller budget without another rollout. Whether such truncated observations faithfully stand in for a native short run has not, to our knowledge, been measured. Comparing them against a same-arm decoding-noise control across budgets from 128 to 2048 output tokens, we find no detectable excess disagreement at budgets of 256 tokens or more, but an excess at 128 tokens (22.5% vs. 11.3% disagreement), the regime where a cost-aware allocator most wants to operate. The excess is bidirectional, so no fixed correction removes it. In an offline replay of an online allocator, reusing truncated feedback helps when tokens are expensive and shows no detectable benefit when they are cheap; removing the 128-token budget from the ladder makes reuse beneficial at the cost weights we tested. Theoretically, we show that pooling biased free observations can be inconsistent, give a regret bound that adapts to which levels' free observations are trusted, and show that each level's free observations saturate, so they cannot replace native pulls. Our results come from one model and 75–160 problems, and no learned policy we tested saves tokens at matched top quality.

## 1 Introduction

Spending more tokens at inference time improves reasoning quality, but the returns are highly uneven across queries: some problems are solved in a hundred tokens and some are not solved in two thousand. That heterogeneity motivates *per-query* compute allocation, and in our own measurements the opportunity is real: a per-query cheapest-correct oracle sits above and to the left of every fixed budget on both math benchmarks (Figure 2 in Appendix A). Against the cheapest highest-accuracy fixed rung it saves $25.4\%$ of tokens *and* gains $2.5$ accuracy points on GSM8K, and saves $42.0\%$ at equal accuracy on MATH-500. How large that gap is depends on the grading instrument, and Appendix A reports it on both.

An online allocator must learn which budget suits which queries, but with bandit feedback it observes only the budget it plays. Truncating one long reasoning trace and scoring each prefix (Figure 1A) yields a *free* observation for every smaller budget without another rollout. The mechanism is not ours (Liao et al., 2025; Qi et al., 2025), and systems that reuse it rest on an assumption we are not aware of anyone having measured: that a truncated prefix is a faithful stand-in for a native short run. If the model conditions its plan on the budget it was given, as budget-guidance methods aim to induce (Li et al., 2025), the free observation is plausibly *biased*, the guarantees of standard graph-feedback analyses (which assume unbiased side observations) do not transfer, and an allocator can be misled rather than accelerated. We measure that bias, trace its consequences for an online allocator, and work out what can still be proved once the free channel is admitted to be biased.

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.

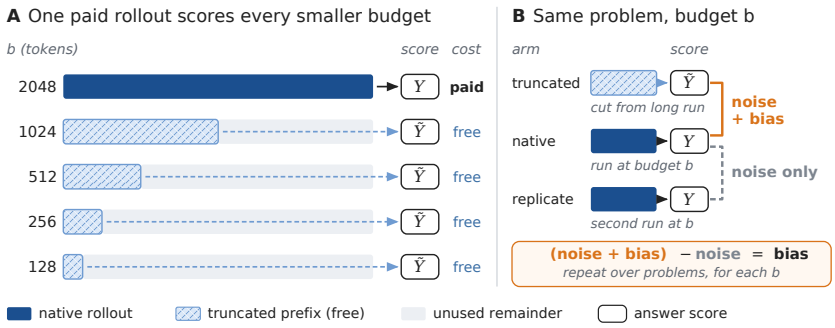


Figure 1: One long rollout yields a score at every smaller budget; we test whether those scores match native runs. (A) A rollout generated once at the largest budget is cut at each smaller budget $b$ and the answer is extracted from each prefix, giving a truncated score $\tilde{Y}$ without any additional rollout; the top row is the paid native score $Y$. (B) For each problem and budget, the truncated score is compared with a native run at $b$ (noise plus bias) and a native run is compared with an independent native replicate (noise only); their difference estimates truncation bias.

**We make three contributions.**

1. **Fidelity of truncated-trace feedback** (§4). Against a same-arm decoding-noise control, truncated-trace feedback is statistically indistinguishable from decoding noise at every budget $\geq 256$ tokens and significantly biased at $b = 128$ (McNemar $p = 0.035$); a mechanism probe finds the disagreement *bidirectional* rather than systematically pessimistic. We establish the *direction* of the $b = 128$ bias, not its magnitude, whose interval spans zero.
2. **What the measured fidelity buys an online allocator** (§5). On a 50-seed offline replay, reusing truncated observations helps at cost weight $\lambda \geq 1$, buys nothing detectable at $\lambda \leq 0.5$, and at $\lambda = 0.25$ is significantly *worse* early. A ladder ablation points to the $b = 128$ rung: with it removed, reuse becomes significantly beneficial at every $\lambda$. A trust threshold that discards that rung changes allocation outcomes, and reverses sign where the distrusted rung is itself optimal.
3. **A deliberately modest theory** (§6). A $\widetilde{O}(\sqrt{T})$ rate already follows from Alon et al. (2015), because the graph is strongly observable with independence number one, and the transitively closed contextual case is treated by Wen et al. (2024); we claim neither. Our new content is soundness under a biased channel: naive pooling can be *inconsistent*, the correct guarantee is an adaptive minimum over per-level trust sets, and a free channel *saturates*: past $n_k^\star = 2\beta_T^2/\delta_k^2$ observations it cannot tighten level $k$'s interval further, so it cannot substitute for native pulls. In simulation at the measured biases the free channel does *not* improve statistical regret; its value is token cost and arm pruning, which we reconcile with the replay result in §6.

**What we do not claim.** We do not claim the formulation of test-time compute allocation as bandit learning (Zuo & Zhu, 2025), a new regret rate, or a deployable win: across six domain–grading cells, *no* learned policy we tested saves tokens at matched top quality, and only hindsight-informed references do. The gain we do find is quality at matched spend (+12.7 points aggregate, +35 on MMLU-Pro), and it appears only where the budget–reward curve is steep and heterogeneous.

## 2 Related Work

**Test-time compute allocation as bandit learning.** Zuo & Zhu (2025) formulate test-time compute allocation as a bandit problem and prove a budget requirement that matches an information-theoretic lower bound. Their arms are *queries*; ours are *budget levels* for a single query, which is the only formulation difference we claim. The ordered discrete budget set is also known: Zhai et al. (2026) order budgets by cost and solve allocation offline by Lagrangian relaxation plus imitation, and Damani et al. (2024) train a probe to predict each budget's marginal reward, both building the

accuracy-versus-budget matrix from independent calls per level. Related controllers pick a decoding strategy under joint token and latency cost, predict a thinking budget from draft agreement, or train models to obey a requested length, alongside work on compute-optimal scaling, efficiency, over- and under-thinking, and models' failure to ration a shared budget on their own (Huang et al., 2025; Lee et al., 2026; Aggarwal & Welleck, 2025; Han et al., 2025; Muennighoff et al., 2025; Snell et al., 2024; Sui et al., 2025; Wang et al., 2025; Fan et al., 2026). This line does not address the online cost of learning the budget–reward profile while serving traffic, and, to our knowledge, does not measure whether a truncated trace can stand in for a native rollout.

**Truncation as a source of multi-budget signal.** Liao et al. (2025) treat truncation depth as a scaling axis, in a cost model whose thinking term carries no factor of the number of depths, and report that truncated traces often match full traces at far fewer tokens. Qi et al. (2025) truncate one thinking process at sampled budgets and force a verifiable answer at each, turning truncation into dense reward for RL. Neither selects a budget online, treats the truncated observations as a feedback structure, analyzes regret, or measures whether the truncated score matches what a natively smaller budget would have earned. That measurement is our lead contribution, and it builds on their mechanism.

**Graph feedback, ordered arms, and multi-fidelity bandits.** Feedback-graph bandits bound regret by the graph's independence number rather than the number of arms (Mannor & Shamir, 2011; Caron et al., 2012; Alon et al., 2015; 2017; Cohen et al., 2016), and ordered-arm bandits exploit an ordering to avoid paying per arm (Yu & Mannor, 2011; Combes & Proutière, 2014; Saber et al., 2021; Varude et al., 2025). Our ladder is strongly observable with independence number one and transitively closed, so a $\widetilde{O}(\sqrt{T})$ rate follows from Alon et al. (2015), and the contextual version of this graph class is treated by Wen et al. (2024) (§6). Since Lykouris et al. (2019) show that graph-blind UCB and Thompson sampling already attain graph-structured rates, we do not present graph-awareness as an advantage.

Three works are closest, and Appendix G gives the precise distinctions. The closest classical model is the multi-fidelity bandit of Kandasamy et al. (2016): ordered cost levels whose cheap fidelities return *biased* estimates. We differ only by *nesting*, which is worth a bounded constant and never a rate (Proposition 6). In Sridhar et al. (2026), the surrogate bias is a per-arm constant and hence estimable; ours has unknown sign and is not identifiable from truncated data (Proposition 13). Agarwal et al. (2023), whose abstract mentions censored feedback on truncated versions of probed variables, study monotonicity over stochastic dominance of $n$ *independent* item distributions with no notion of context, whereas our $K$ levels are dependent views of one latent trace in a contextual setting. Further context (pure-exploration budget ladders, cost-constrained and contextual bandits, bandits for LLM routing that use model identity as the arm, and quality-scaling methods) is in Appendix G.

## 3 SETUP

**Budget ladder and truncation feedback.** Fix one generator (throughout, a single hosted frontier reasoning model, with the exact model identifier and the separate extractor model recorded per row in the released measurement table) and an increasing ladder of decoding budgets $b_1 < \cdots < b_K$; throughout, $\mathcal{B} = \{128, 256, 512, 1024, 2048\}$. *Budget here is a hard cap on output tokens, covering reasoning and answer together*, because the model we use exposes no native thinking-budget parameter. This is a real limitation, since our ladder is not a thinking allocation. Let $\tau_j(\cdot)$ truncate a generated sequence to its first $b_j$ tokens and then run the same answer extraction and grader used everywhere else (Figure 1A). Playing level $k$ produces one rollout $\rho$ and, for every $j \le k$, a truncated score $\widetilde{Y}(j) = \mathrm{score}(\tau_j(\rho))$ at no additional generation cost; we call $\widetilde{Y}(j)$ for $j < k$ the *truncated channel* and the score of the untruncated rollout the *native channel*. The induced feedback graph has $k \to j$ for all $j \le k$: complete lower-triangular, transitively closed. As §1 states, this observation structure is a reading of Liao et al. (2025) and Qi et al. (2025), not a discovery of ours.

**Utility.** With known costs $c_k$ normalized to $b_k/b_K$, per-query quality $q(x,k) \in [0,1]$ and a deployer-chosen cost weight $\lambda \ge 0$, the utility of level $k$ is $u_\lambda(x,k) = q(x,k) - \lambda c_k$, and regret is measured against the *per-context* optimum $\max_k u_\lambda(x,k)$ rather than the best fixed

level. Empirically, cost is in kilotokens of measured completion tokens and $\lambda$ is swept over $\{0, 0.1, 0.25, 0.5, 1, 2, 4\}$. **Every conclusion here is $\lambda$-dependent and several change sign inside that grid**; a reader with a specific operating point should read the row, not the summary.

**What is assumed.** Appendix I states four assumptions and records what our data says about each; two matter here. *Truncation fidelity* posits a per-level, **two-sided** bias bound $\delta_j$, which §4 measures rather than assumes, and whose two-sidedness is forced by the measurement rather than chosen for convenience. *Monotonicity of quality in budget* is supported only in the weak in-expectation form and contradicted per instance ($17.5\%$ of GSM8K and $8.6\%$ of MATH-500 problems flip from correct to wrong as the budget grows): **no result in this paper uses monotonicity**, it drives only an optional accelerator, and Proposition 14 exhibits an algorithm that fails catastrophically for relying on it.

**Benchmarks, grading, and two evaluation universes.** We use GSM8K (Cobbe et al., 2021), MATH-500 (the 500-problem subset of MATH (Hendrycks et al., 2021) used by Lightman et al. (2024)), MBPP (Austin et al., 2021), and MMLU-Pro (Wang et al., 2024). Grading instruments differ by necessity, and rather than mix them or impute missing cells we keep two strictly separated universes. PRIMARY is GSM8K + MATH-500 graded by forced-answer extraction ($n = 75$ problems with a complete two-replicate ladder); it is the only universe in which the truncated arm was graded, so every truncation-reuse result is confined to it. BREADTH is all four domains graded by a self-reported answer line, with MBPP graded by executing the dataset's own unit tests ($n = 160$), and has no truncation-feedback variant. **MBPP and MMLU-Pro forced-answer accuracy, and MBPP/MMLU-Pro triage features, are unmeasured, never zero**: those rows are excluded by domain, never imputed, and a hard assertion fails if any missing value reaches a design matrix (Appendix B).

## 4 HOW FAITHFUL IS A TRUNCATED TRACE?

**Design: three quantities, one grader, a same-arm control.** For each target budget $b_j$ we measure three quantities on the *same* problems through the *identical* extraction and grading path, so the only thing that varies is where the prefix came from: (a) NOISE, the disagreement between two *native* $b_j$ rollouts (replicate A vs. B), the decoding-noise floor; (b) TOTAL, the disagreement between a rollout truncated $2048 \to b_j$ and a native $b_j$ rollout, i.e. noise plus truncation bias; and (c) BIAS = (b) − (a), the truncation bias $\widehat{\delta_j}$ (Figure 1B). Both series are structurally matched — in each the two compared rewards come from two separate sampling events — which is what makes (a) the correct null for (b), and without (a) an agreement rate is uninterpretable: our own replicate disagreement runs from $0\%$ to $25\%$ across domains and budgets (Appendix C).

**Why McNemar, and the internal calibration control.** The two series share problems and the same native-B rollout, so the per-problem indicators "did this comparison agree?" are *paired*; a two-proportion test would treat them as independent, overstate the variance and lose power, so we use exact binomial McNemar on discordant pairs with Wilson intervals throughout. At $b = 2048$ the "truncated" arm *is* the untruncated trace, so its disagreement with a native rollout isolates extractor stochasticity alone: $1/80 = 1.25\%$ (Wilson CI 0.2–6.8%). The measured "bias" there is $0.013$ — it lands on that floor, as it must if the decomposition is calibrated, and the $0.013$–$0.038$ values elsewhere on the ladder sit inside the same floor.

**Result: faithful above 256 tokens, biased at 128.** Table 1 gives the result, plotted in Figure 3 in Appendix A. **Truncated-trace feedback is statistically indistinguishable from decoding noise at every budget except the shortest.** At $b \in \{256, 1024, 2048\}$ the excess disagreement is 1.2–1.3 points, the size of the extractor-only control, with McNemar $p = 1.0$; at $b = 512$ it is 3.8 points, $p = 0.45$. At $b = 128$ **the bias is real**: $22.5\%$ versus $11.3\%$ disagreement, an excess of 11.3 points, $p = 0.035$, from 12 problems where two native rollouts agreed but the truncated arm disagreed against 3 the other way.

So the field's implicit assumption holds over most of the ladder and breaks at the shortest budget — which is exactly where a cost-aware allocator wants to operate. That is the modelling term an

Table 1: Truncation fidelity against a same-arm decoding-noise control. Forced-answer grading, exact binomial McNemar on paired per-problem agreement; $\widehat{\delta_j}$ is the excess disagreement and $b/c$ the discordant-pair counts. The $b = 2048$ row is the calibration control (there the "truncated" arm is the full trace) and its $0.013$ is the extractor floor.

| $b_j$ | $n$ | (a) noise | (b) total | $\widehat{\delta_j}$ | 95% CI on $\widehat{\delta_j}$ | $b$ | $c$ | McNemar $p$ |
|---|---|---|---|---|---|---|---|---|
| 128 | 80 | 0.1125 | 0.2250 | **0.1125** | $[-0.053, +0.268]$ | 12 | 3 | **0.035** |
| 256 | 80 | 0.0875 | 0.1000 | 0.0125 | $[-0.118, +0.142]$ | 5 | 4 | 1.000 |
| 512 | 79 | 0.1013 | 0.1392 | 0.0380 | $[-0.108, +0.180]$ | 5 | 2 | 0.453 |
| 1024 | 80 | 0.0375 | 0.0500 | 0.0125 | $[-0.085, +0.109]$ | 4 | 3 | 1.000 |
| 2048 | 77 | 0.0779 | 0.0909 | 0.0130 | $[-0.115, +0.140]$ | 1 | 0 | 1.000 |

algorithm must carry: a per-level feedback-bias parameter that is near zero for $b \geq 256$ and about $0.11$ at $b = 128$.

**Direction established, magnitude not.** We state this precisely because it is the sharpest thing a reviewer can press on. The paired McNemar test rejects at $b = 128$, but the interval on the excess-disagreement *magnitude* is $[-5.3, +26.8]$ points and spans zero. The two are consistent — McNemar conditions on discordant pairs and is better powered at $n = 80$ — and the honest statement is that *the direction of the $b = 128$ bias is established, its magnitude is not*. Everything downstream uses the direction only.

**Mechanism: the disagreement is bidirectional, not pessimistic.** The intuitive failure mode is that truncation is systematically pessimistic: a trace cut mid-derivation has no answer to extract, whereas a natively short rollout would have been forced to conclude. **Our data rejects that model.** Classifying all 41 disagreeing pairs from observable signals gives 22 cases where the truncated arm is correct and the native arm wrong, 17 the reverse (10 for reasons other than truncation, 7 where the native rollout compressed its reasoning to fit), and only **2** cut mid-derivation with nothing extractable.

Disagreement is therefore near-symmetric and slightly favours the truncated arm (22 against 17): short native rollouts fail about as often as truncated prefixes, frequently by not committing to an answer inside the cap. Two consequences bind the rest of the paper — the bias bound must be stated **two-sided** with no sign assumed, and there is no constant offset to remove, so de-biasing is unavailable, which Proposition 13 later shows is an identifiability failure rather than a practical inconvenience. The probe is rule-based, not human- or LLM-adjudicated, and we do not present it as a validated taxonomy. Directionality in the source–target gap is **unmeasured**: the $b = 128$ result is consistent with bias growing in $k - j$ (the largest gap tested, $2048 \to 128$, is the only significant cell) but only one source budget was tested, so we flag this as suggestive rather than established.

## 5 What the fidelity buys an online allocator

**Protocol.** Because the measurement recorded the graded reward of *every* rung for *every* problem, any policy that picks one rung per problem can be scored exactly against the measured reward and token cost of the rung it picked. This is an **offline replay**, not a deployment: the recorded reward of an arm is taken to be what the policy would have obtained, and replay cannot capture any effect the policy's own choice would have on the generation. Problems are drawn with replacement in a seed-specific shuffled order, $T = 750$ per stream, **50 seeds**; on each visit one of the two measured replicates is drawn uniformly, so the realized reward is a genuine second sample, not a table lookup. Every comparison is paired within seed with a 95% CI over seeds, $p$-values from paired $t$-tests corroborated by Wilcoxon signed-rank, Holm-corrected across the seven-$\lambda$ family per question. Exploration rates were tuned on disjoint seeds and frozen at one shared value; policies reading a triage feature are charged its output tokens at parity with generator tokens, conservative against them (Appendix E).

**Q1: does reusing truncated feedback buy anything? Yes when tokens are expensive; nothing detectable when they are cheap.** Table 2 (left) gives the paired per-seed utility advantage of reusing

Table 2: Paired per-seed utility deltas, 50 seeds, PRIMARY universe, Holm-corrected within each question. **Q1:** reusing truncated observations minus bandit feedback only. **Q2:** the trust threshold (discard the $b$=128 truncated observations) minus naive reuse of all levels. Bold marks Holm $p < 0.05$; every entry without bold has a 95% CI spanning zero. Q1 is null at low $\lambda$ with three of four LinUCB point estimates slightly negative; Q2 *reverses sign* at high $\lambda$. Table 6 in the appendix gives every interval and $p$-value.

| | Q1: reuse − bandit-only | | Q2: trust − naive reuse | |
|---|---|---|---|---|
| $\lambda$ | LinUCB | LinTS | LinUCB | LinTS |
| 0 | −0.0033 | −0.0076 | +0.0070 | +0.0085 |
| 0.1 | −0.0048 | −0.0045 | +0.0082 | **+0.0126** |
| 0.25 | −0.0054 | −0.0002 | **+0.0130** | **+0.0111** |
| 0.5 | +0.0040 | +0.0046 | **+0.0117** | **+0.0128** |
| 1.0 | **+0.0189** | **+0.0160** | **+0.0076** | **+0.0095** |
| 2.0 | **+0.0341** | **+0.0426** | **−0.0089** | −0.0048 |
| 4.0 | **+0.0384** | **+0.0420** | **−0.0185** | **−0.0184** |

truncated observations over bandit feedback alone; Figure 5 in the appendix plots the regret curves and the trust-threshold sign reversal.

At $\lambda \geq 1$ both learners gain $+0.016$ to $+0.043$ utility with Holm $p < 0.005$, and the mechanism explains the $\lambda$-dependence: at low $\lambda$ the optimal rung is expensive and free information about cheap rungs is information the learner does not need, whereas at high $\lambda$ the optimal rung is cheap, a bandit-feedback learner must pay full price to learn about cheap rungs, and truncation delivers exactly that for free.

**Negative result, stated as such:** at $\lambda \leq 0.5$ the feedback is free and buys nothing detectable at 50 seeds — every CI spans zero and three of the four low-$\lambda$ LinUCB point estimates are slightly negative. We claim neither harm nor benefit there. On learning speed the picture is sharper. Cumulative regret against the per-query oracle (reuse minus bandit-only) at $\lambda = 2.0$ is $-3.30$ $[-4.48, -2.12]$ at 50 problems, $-7.36$ at 200 and $-25.57$ at 750 — the advantage appears immediately and compounds. At $\lambda = 0.25$ it does not merely vanish: **reuse is significantly worse early**, $+0.94$ $[+0.30, +1.57]$ excess regret at 50 problems ($p = 0.006$), converging to no detectable difference by 750. Reusing observations the measurement showed to be biased at the shortest rung actively misleads a learner whose optimum lies at a longer rung, and it takes a few hundred problems to recover.

**Q2: does the measured $b$=128 bias matter in practice? Yes at moderate cost weight, and the sign reverses at extreme cost weight.** Table 2 (right) shows the trust threshold — reuse truncated observations only for levels $\geq 256$, motivated directly by §4 — beating naive reuse at $\lambda \in [0.25, 1]$ for both learners and *losing* at $\lambda \geq 2$. The reason is not subtle: at $\lambda \geq 2$ the optimal rung *is* $b = 128$, so discarding its observations discards exactly what the learner most needs, and that information starvation costs more than the bias does. The measured bias is worth respecting where the allocator has alternatives; a blanket "never trust the shortest rung" rule is wrong when the shortest rung is the answer. Against the bandit-feedback baseline the threshold is positive at every $\lambda$ for both learners (LinUCB $+0.0035$ to $+0.0265$), significant after correction at $\lambda = 0.5$–$4.0$; at $\lambda = 0.25$ the learners straddle the conventional threshold (LinTS Holm $p = 0.029$, LinUCB $p = 0.0515$) so that cell is suggestive only, and we claim no significance at $\lambda \leq 0.1$.

**The ladder ablation is the decisive evidence.** If the $b$=128 bias is what damages naive reuse, deleting $b$=128 from the ladder should make reuse uniformly good — and it does. Q1 deltas at $\lambda = (0.25, 0.5, 1.0, 2.0)$: full ladder $(-0.0054, +0.0040, +0.0189, +0.0341)$ and $\{128, 512, 2048\}$ $(-0.0206, -0.0047, +0.0149, +0.0322)$, against $\{256, 512, 1024\}$ $(+0.0146, +0.0216, +0.0353, +0.0384)$ and $\{256, 2048\}$ $(+0.0170, +0.0251, +0.0292, +0.0234)$. **Both $b$=128-free ladders benefit significantly at *every* $\lambda$, including the low ones where the full ladder showed nothing (all $p < 0.0001$); both $b$=128-containing ladders show the same low-$\lambda$ null or negative.** This is an independent line of evidence that the single measured biased level erases the low-$\lambda$ benefit, and it is the strongest

practical form of the fidelity result: the measurement told us which level to distrust, and acting on it changed the outcome.

**Q3: how much oracle headroom is captured, and at what cost?** The honest headline is a negative result with a sharp boundary. Because domains disagree about the right rung (at $\lambda = 0.5$ the $\lambda$-optimal fixed rung is 512 for GSM8K, 1024 for MATH-500, 2048 for MBPP and MMLU-Pro), a policy can beat one global fixed budget by identifying the domain and nothing more, so we report headroom against *both* that and the **best fixed rung per domain chosen in hindsight** (Table 7). **In the two-math-domain PRIMARY universe no learned policy beats the best fixed budget**: its best rung ($b = 256$) is already near-optimal for both domains, the oracle gap is small (0.11–0.15 utility), triage side-cost plus estimation error consume more than the gap, and every learned policy captures a *negative* fraction — we report the universe where our own method loses. In the four-domain BREADTH universe the supervised router captures 31–53% of the headroom at $\lambda \geq 0.5$ and beats the per-domain hindsight reference (0.308 vs. 0.067 at $\lambda = 0.5$), so it is doing genuine per-query allocation rather than domain identification — within MATH-500 at $\lambda = 0.5$ it splits $12/10/0/30/48\%$ across the ladder rather than committing to a rung. Per domain at $\lambda = 1.0$ it captures $+0.744$ $[+0.677, +0.811]$ of MMLU-Pro's gap and $+0.382$ $[+0.251, +0.513]$ of MBPP's, but $+0.075$ and $+0.026$ (CIs spanning zero) on self-report MATH-500 and GSM8K and $-0.920$ on both forced-answer math cells: **per-query allocation helps where the budget–reward curve is steep and heterogeneous and not on the two math domains at all.**

**No learned policy saves tokens at matched top quality.** Taking the quality of the highest-quality fixed budget as the target, the oracle saves 11–64% across the six domain–grading cells and no learned policy reaches the target with a saving on any of them (Table 9). Two references do save — the cheapest-correct oracle (8.9% on PRIMARY GSM8K) and the per-domain best fixed rung (11.6% on the PRIMARY aggregate) — but both need hindsight a deployed system does not have, so the accurate statement is: *only hindsight-informed references save tokens at matched top quality; nothing fitted from decision-time features does.* What per-query allocation does buy is **quality at matched spend**: against the fixed-budget curve interpolated at the same token spend the router gains $+0.353$ on BREADTH MMLU-Pro (0.660 vs. 0.307 at 738 tokens), $+0.127$ aggregate, and only $+0.020$ and $+0.012$ on the two PRIMARY cells — and MMLU-Pro, where it buys the most, is also the domain whose headroom is partly a failure-to-terminate artifact (§7). Appendix F gives the feature, triage-price and cross-domain ablations; their upshot is to use the free structural features and skip the triage call at any meaningful cost weight, and that the decision-time features do not transfer even between two math datasets.

## 6 THEORY: WHAT A BIASED FREE CHANNEL CHANGES

Statements and framing only; the formalization, the algorithm TRUNCUCB, and every proof are in Appendices I–L.

**What we concede.** The nested feedback graph ($k \to j$ for all $j \leq k$) is strongly observable with independence number $\alpha = 1$, domination number 1, and is transitively closed (Proposition 5), so by the trichotomy of Alon et al. (2015) the minimax regret is $\widetilde{\Theta}(\sqrt{\alpha T}) = \widetilde{\Theta}(\sqrt{T})$, independent of $K$. **We claim no credit for a $\widetilde{O}(\sqrt{T})$ rate on compute ladders; any such bound, including the one our own algorithm satisfies, is a corollary of a 2015 theorem.** The transitively closed *contextual* case is covered by Wen et al. (2024), and since Lykouris et al. (2019) show graph-blind UCB and Thompson sampling already attain graph-structured rates, **"our algorithm uses the feedback graph" is not itself a contribution** and our ablations must not be read as one. Nesting — the one structural feature separating compute ladders from the multi-fidelity bandit of Kandasamy et al. (2016) — is worth only a bounded constant: covering the ladder costs $b_K$ tokens nested against $\sum_k b_k$ paid per level, a saving of $\rho_K = 1.94$ here and at most $r/(r-1)$ for any geometric ladder however finely subdivided (Proposition 6), so *any claim that nesting improves the **rate** over MF-UCB would be false*.

**The gap: the generic machinery is unsound here.** Graph-feedback analyses assume side observations are unbiased draws of the neighbour's reward. §4 shows they are not, and the consequence is an inconsistent algorithm, not a worse constant. Any algorithm that pools truncated and native

observations of a level into one unweighted least-squares estimator — the literal reading of the $\alpha = 1$ reduction — has regret $\Omega(\delta_j T)$, i.e. *linear* (Proposition 7): the pooled estimator converges to a displaced limit, so a level whose true utility advantage is smaller than the bias is ranked incorrectly forever however much data arrives. Simulation reproduces it: the naive learner's log-log regret slope rises from $0.45$ at zero bias to $0.53$ at $8\times$ the measured bias while its regret grows from 554 to 704. **Invoking $\alpha = 1$ on a compute ladder is unsound, not merely loose.** TRUNCUCB therefore keeps two estimators per level and never mixes their samples: a native channel, unbiased but fed at most one sample per round, and a pooled channel, fed $m_k = \sum_{i \geq k} n_i$ samples but biased by at most $\delta_k$ with *unknown sign*, so valid only after *symmetric* inflation. Theorem 8 then gives a regret bound that is an *adaptive minimum* over trust sets of the levels whose pooled channel is admitted, and is $\widetilde{O}(\beta_T \sqrt{dKT})$ always since the empty trust set is in the minimum. **No monotonicity assumption is used**, deliberately: an algorithm enforcing $\varepsilon$-monotonicity by capping lower levels' bounds via higher ones drives the optimal level's interval *empty* and incurs $\mathbf{22.4}\times$ the regret of the monotonicity-free default, playing the costliest level on $65.1\%$ of rounds (Proposition 14) — not hypothetical, given measured violation rates of $17.5\%$ and $8.6\%$. Theorem 12 separately prunes cost-dominated arms with no data and no assumptions, replacing $K$ by a window size $W$.

**Saturation is the real contribution.** Level $k$'s pooled channel is worth admitting precisely when $\delta_k \lesssim \tau_T = \sigma \beta_T \sqrt{dKL/T}$, and it stops improving once it holds $n_k \gtrsim n_k^\star = 2\beta_T^2/\delta_k^2$ observations, beyond which its interval is pinned at the bias floor $\delta_k \sqrt{2}/\sigma$. The reason is exact cancellation: the pooled channel's stochastic width shrinks as $n_k^{-1/2}$ while its bias displacement is $\delta_k \sqrt{n_k} \|\phi\|_{V^{-1}}$ with $\|\phi\|_{V^{-1}} \asymp \sqrt{2/(\sigma^2 n_k)}$ under context diversity, so the $\sqrt{n_k}$ cancels and the bias contribution is *constant* in $n_k$ (Lemma 18). **Free observations buy precision only up to $n_k^\star$ and nothing after, so the free channel cannot substitute for native pulls** — a distinction invisible to $\alpha$-based analysis. At the measured $\delta_k$ ($\beta_T = 4.90$, $T = 4 \times 10^4$, $\sigma^2 = 0.098$ measured), $n^\star = 3.8 \times 10^3$ at $b$=128 against $3.3 \times 10^4$ at $b$=512 and $\approx 3 \times 10^5$ elsewhere — nearly two orders of magnitude separating the one biased rung from the rest (Remark 10). The floor is *information-theoretic*, not an analysis artifact: two instances — quality $v$ with zero bias, and quality $v + \delta_j$ with bias $-\delta_j$ — induce identical laws for every truncated observation (joint TV distance 0, verified by enumeration) yet have different optimal levels, so no truncation-only algorithm can separate them and native pulls are necessary (Proposition 13). This also distinguishes our setting from Kandasamy et al. (2016), whose fidelity gaps are known, and Sridhar et al. (2026), whose surrogate bias is estimable.

**Theory and measurement pick out the same rung, with an exact caveat.** Since $\tau_T$ strictly decreases in $T$, **levels leave the trust set in decreasing order of $\delta_k$ and the first to leave is the rung with the largest measured bias** — at the measured $\widehat{\delta}$, $b$=128, then $b$=512, then the rest. The measurement, performed independently, finds $b$=128 the *only* rung whose bias is distinguishable from decoding noise ($p = 0.035$; all others $p \geq 0.45$) — two unrelated routes select the same rung, and **no constant was tuned to make $b$=128 come out first**. The caveat is equally exact: the *ordering* is derived from the bound with no experimental input beyond the $\widehat{\delta}_k$, whereas the *absolute* threshold depends on constants we fixed rather than derived ($\varrho = 1$, confidence $0.05$, the $O(1)$ in $\beta_T$), so at our verification horizon $T = 4 \times 10^4$, $\tau_T = 0.125 > 0.1125$ and *every* level is still admitted, $b$=128 excluded only for $T \gtrsim 5.2 \times 10^4$. We claim the ordering and the first rung dropped, not a horizon-free cutoff.

**The honest negative result, and why it does not contradict §5. At the measured $\delta_k$ the free channel does not improve regret.** On 24 paired instances at $T = 2 \times 10^4$, TRUNCUCB is slightly *worse* than the native channel alone: $-18.4 \pm 6.3$ regret in its disfavour, $p = 0.0033$ *against* it (per-seed spread, sd $\approx 60$ on a mean $\approx 500$, exceeds the between-algorithm gap, hence a paired test). An earlier draft claimed a 14–19% improvement, an artifact of a one-sided bias model that vanished once the model matched §4. Relatedly, **the trust rule is not empirically separable at the measured biases** — gated and ungated variants are statistically indistinguishable, and the rule bites only when biases are several times larger, so it is a safeguard rather than a demonstrated simulation win. A reviewer will notice an apparent tension with §5, where reuse *does* pay ($+0.019$ to $+0.038$ utility at $\lambda \geq 1$, Holm $p < 0.001$), so we state the resolution rather than leave it to be found: **they measure different things.** The simulation measures pure statistical regret, where extra observations help only by narrowing a level's interval — which saturation says the biased ones stop doing. The

replay measures utility under a *cost-weighted* objective, where at high $\lambda$ the optimal rung is cheap and what matters is information about *cheap* rungs, which a bandit-feedback learner can get only by paying for them and truncation supplies at zero generation cost. The synthesis is therefore narrower than "truncated feedback helps": it is faithful enough to be *sound* at $b \geq 256$, and its value lies in **token cost** (the $\rho_K$ nesting constant) and cost-domination pruning, **not in statistical efficiency**. Reusing it naively is unsound in theory and, at these bias levels, does not pay statistically. Table 10 reports the numerical verification, including the rows whose prediction is a negative one.

## 7 LIMITATIONS

**Scale, and what the intervals cover.** The fidelity measurement rests on $n \approx 80$ problems per budget and 40 per domain for the noise floor; the replay universes have $n = 75$ and $n = 160$, with 35–40 problems per domain. The 50-seed intervals in §5 cover **replay streams, not problem resampling**: they control stream variance and say nothing about what a different problem sample would give, so problem-level uncertainty is substantially larger. **Single model, and a cap rather than a thinking budget.** All measurements come from one generator and one extractor, so we have not tested whether the $b = 128$ bias is a property of this model or of budget-conditioned reasoning generally; the budget is a hard output-token cap, not a native thinking allocation; and we ran replicates at provider-default decoding (temperature is deprecated for this model), so the noise floor is the floor *at default decoding*. **The $b = 128$ bias: direction yes, magnitude no.** McNemar rejects ($p = 0.035$) while the interval on the excess-disagreement magnitude spans zero ($[-5.3, +26.8]$); everything we build on it uses the direction only. **Replay, not deployment.** §5 scores policies against the recorded reward of the rung picked, so it cannot capture interaction between a policy's choice and the generation, and we ran no live serving experiment. **Unmeasured cells restrict scope rather than being imputed.** We did not measure MBPP and MMLU-Pro forced-answer accuracy or their triage features, and we do not treat them as zero; we graded the truncated arm only on the forced-answer path, so Q1 and Q2 are two-domain results, and MATH-500 forced-answer coverage is partial. The two grading instruments give different oracle-versus-fixed gaps on the same problems, so every headroom number names its path. **MMLU-Pro's headroom is partly an artifact**: its accuracy climbs from $2.5\%$ to $77.5\%$ largely because the model does not emit an answer line within short caps rather than because it reasons wrongly, so the domain where per-query allocation helps most is also where the metric is most sensitive to answer formatting. It also substitutes for an unreachable personalization benchmark, so nothing here speaks to user-conditioned reward. **Truncation granularity and the directionality gap.** Cut points are approximate to within a token or two, and we did not measure bias versus the source–target gap $k - j$, because we used one source budget only. **Theory gaps.** We assume realizability of a per-level linear quality model and did not test it, and context diversity remains open (the native branch never needs it); we prove *no* lower bound matching the adaptive minimum; the monotone-capping counterexample is a verified failure on an instance family rather than a closed-form $\Omega(T)$; our bias lemma gives magnitude without sign, so no one-sided sharpening of the pooled interval is available; and our constants are not sharp. Appendix H expands every item above and lists the next measurements: bias versus the truncation gap, a second model, and a native thinking-budget parameter.

## 8 CONCLUSION

We measured how faithfully a truncated reasoning trace stands in for a native run at a smaller budget. Against a same-arm decoding-noise control, we found no detectable bias at budgets of 256 tokens or more, but a bias at 128 tokens, where a cost-aware allocator most wants to operate. Because the bias is bidirectional, no fixed correction removes it. Our replay and theory support a narrower conclusion than "truncated feedback helps": it is sound above the shortest rung, its value lies in token cost and arm pruning rather than statistical efficiency, and it cannot substitute for native pulls. Our results come from one model, and testing a second model, the dependence of the bias on the truncation gap, and a native thinking-budget parameter are the natural next steps.

## REPRODUCIBILITY STATEMENT

Every number in this paper is read out of a saved table, and the tables are produced by scripts we will release.

**Measurement (§4).** The generation harness records one row per (problem, budget, arm type, replicate, grading path) with text length, actual completion and prompt tokens, per-call latency, stop reason, graded reward, grade kind, the exact generator and extractor model identifiers, and the decision-time context features — 2910 rows over four domains. The fidelity decomposition, the exact-binomial McNemar tests, the Wilson intervals, the extractor-only control at $b = 2048$, the mechanism-probe classification, and the four-domain noise floor are each computed from that table by a single analysis script. Truncation is character-based with per-trace chars-per-token calibration, and its parameters are recorded per row so cut points are reproducible.

**Allocation replay (§5).** The replay is fully offline and requires no API access: universe construction and coverage rules, all policies, the 50-seed sweep, the ablations and the analysis are separate scripts, deterministic given the published seed list. Exploration rates were tuned on seeds 200–211 and frozen; those seeds appear in no reported result. Held-out scoring uses a disjoint 50% problem split.

**Theory (§6).** Appendix L contains complete proofs of every stated result. The verification script is runnable and reproduces Table 10; it deliberately uses the theoretical confidence radius with no tuned exploration constant, so "empirical $\leq$ bound" is a real test rather than a fit, and it measures $\sigma^2$ as $\lambda_{\min}(\mathbb{E}[\phi\phi^\top])$ from the design rather than assuming it.

**Three provenance disclosures.** These are exactly what a reproducibility statement is for, so we state them plainly rather than in a footnote. (i) *The theory verification script was destroyed and rebuilt mid-project.* A string-replacement bug destroyed `verify_theory.py`; it was rebuilt, and every theory number in this paper comes from a *single run of the rebuilt script*, each value checked programmatically against the saved summary JSON, with no value carried over from the pre-crash run. The rebuild reproduces every qualitative conclusion but changed 104 field values non-trivially — log-log slope estimates differ by up to $\approx 25\%$ and the free-channel-versus-native difference flipped sign ($+0.010$ to $-0.027$), that last being precisely the quantity too small to be significant, which is why it is reported through a paired test rather than as a point estimate. RNG consumption order changed in the rebuild, so per-seed trajectories differ. (ii) *The headroom estimator was revised after an audit.* It was changed from a mean of per-seed ratios to a ratio of means after 148 non-finite per-seed ratios were found, arising from streams with zero oracle gap; the revision shifts every headroom number by 0.01–0.05 and reverses no conclusion. (iii) *741 measurement records were quarantined, not used.* They were infrastructure failures that an early recorder had written as reward-0 data. The harness now classifies infrastructure failure as unmeasured and refuses to persist it; the quarantined file is retained for audit.

**Other data-quality controls that were actually run**, as distinct from intended: the replay substrate asserts a finite reward and finite token cost for every retained cell, so an incomplete ladder cannot enter a mean; the feature builder asserts that no missing value reaches a design matrix, and this assertion fired during development and was resolved by restricting the feature set rather than by filling values; and every number in the underlying reports was re-read from the saved tables and diffed against the prose, which caught and fixed two errors (a column misread in one headroom row, and regret figures left from an earlier 30-seed run).

**What cannot be reproduced exactly.** Generations come from a hosted model API at provider-default decoding (temperature is deprecated for this model), so re-running the generation stage will not reproduce the traces token-for-token. The replay stage, all statistics, and all figures reproduce exactly from the released measurement table.

## AI USE STATEMENT

This disclosure is reconstructed from a contemporaneous task log kept while the work was carried out, and it is deliberately expansive rather than minimal. **This paper was drafted primarily by an AI agent under human direction, and generative AI was used at every stage of the project.** The categories below follow the ICLR 2027 AI Policy for Authors, retrieved from the conference site rather than recalled.

**Tasks with required disclosure, for which we used generative AI.** *Propose or refine hypotheses*: the central hypothesis — that a trace truncated from a larger-budget rollout is a biased rather than free stand-in for a native small-budget rollout, because the model plans against the budget it is given — was AI-formulated, as was the re-scoping of the paper after a literature audit found the original angle occupied. *Design or provide feedback on research methodology or experiments*: the three-way fidelity decomposition, the choice of the exact binomial McNemar test over a two-proportion test and the reason for it, the $b = 2048$ self-calibration control, the two-universe separation that keeps grading instruments from being mixed, the offline-replay protocol with tuning seeds disjoint from evaluation seeds, the dual regret reference including the hindsight per-domain fixed rung, and the ladder ablation were all AI-designed. *Help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs*: the formalization and its named assumptions, the saturation theorem, the inconsistency result for naive pooling, the nesting-constant result, the information-theoretic identifiability floor, the monotonicity counterexample, the cost-domination theorem, and all proofs were AI-produced. *Implement methods*: the measurement harness, the unit-test-execution grader, all allocation policies, the replay driver, the algorithm TRUNCUCB, and the numerical verification were AI-implemented. *Clean and reformat dataset*: benchmark loading, answer extraction, the grading paths, the coverage discipline that blanks rather than zeroes unmeasured cells, and the quarantine of 741 records in which API failures had been recorded as reward-0 observations. *Support qualitative and thematic data analysis*: the literature scan classified several hundred papers along a structured schema using LLM calls; the mechanism probe of §4 is rule-based, and we say so rather than presenting it as adjudicated. *Interpret results*: the readings throughout §§4–6, including the explanation of the cost-weight dependence, the trust threshold's sign reversal, and the reconciliation between the simulation's negative statistical finding and the replay's positive utility finding, are AI-proposed and human-reviewed against the saved tables.

**Required categories that do not apply.** We did not generate synthetic data sets for training or evaluation — the only simulated data are the calibrated instances used to numerically verify the theory, which are described as such in §6 and Table 10 — and we did not use AI for translation. No research material was transcribed from recordings.

**Tasks with recommended disclosure, for which we used generative AI.** Sourcing and searching for information, including retrieving the conference's own policy pages; identifying relevant literature and summarizing or analysing it (two independent scans, the second deliberately un-anchored on the first, with full texts pulled for the closest works); discovering research topics and identifying gaps; brainstorming; creating and editing all software code; creating and modifying every figure and table; creation of artifacts; drafting all parts of this paper; editing for readability; suggesting its structure; formatting references; and proposing its title.

**How the AI-assisted work was reviewed, and what review found.** We have reviewed all AI-assisted work, and we record the failures because a disclosure implying reliability would itself be inaccurate. Every number was re-read from a saved table and diffed against the prose; that process caught a headroom row read from the wrong column and stale regret figures from an earlier seed count, and the corrected values revealed an additional finding (that reuse is significantly worse early at $\lambda = 0.25$) that the stale numbers had hidden. An overstated significance claim and an overstated "no policy saves tokens" claim were both narrowed after audit. On the theory side, a first simulation harness had miscalibrated confidence radii and was rewritten; an initial proof step asserted a sign for the bias displacement that the lemma does not establish and was corrected to a symmetric inflation; a reported $p$-value disagreed with the value the script computed and was corrected with both estimators named; and **an earlier draft claimed a $14$–$19\%$ regret improvement from the free channel that was an artifact of a one-sided bias model — once the model matched the measurement the effect reversed, and it is now reported as a negative result in §6**. A statement that all bibliography entries had been verified was written before two had actually been checked; the checks were then performed and found a missing author. Bibliography entries were each verified against arXiv, Crossref, or OpenAlex, which corrected three author lists and excluded one commonly miscited arXiv identifier that points to an unrelated paper. The theory verification script was destroyed by a string-replacement bug and rebuilt, which changed 104 reported field values; this is disclosed in the Reproducibility Statement. The simulation was never adjusted to fit a bound; where it contradicted

a draft of the theory, the theory was corrected. Human decisions in this project were the approval of the plan, the choice of the narrower empirical-led scope over shelving the work, and the acceptance of the mixed and partly negative results as the paper's actual finding.

We take responsibility for the final content of this work, including all text, claims, code, figures, and proofs produced with the aid of generative AI.

## REFERENCES


Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011.

Arpit Agarwal, Rohan Ghuge, Viswanath Nagarajan, and Zhengjia Zhuo. Semi-bandit learning for monotone stochastic optimization, 2023. arXiv preprint arXiv:2312.15427.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning, 2025. arXiv:2503.04697.

Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks, 2016. NeurIPS 2016; arXiv:1507.06738.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 127–135, 2013. arXiv:1209.3352.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, 2015. arXiv:1502.07617.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017. doi: 10.1137/140989455.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. arXiv:2108.07732.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 207–216, 2013. doi: 10.1109/FOCS.2013.30.

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

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits, 2012. UAI 2012; arXiv:1210.4839.

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

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards scaling laws of compound inference systems, 2024. NeurIPS 2024; arXiv:2403.02419.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

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

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs, 2016. ICML 2016; arXiv:1605.07018.

Richard Combes and Alexandre Proutière. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014. arXiv:1405.5096.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of lm computation, 2024. arXiv preprint arXiv:2410.04707v1.

Rodrigo Guedes de Souza and Alison R. Panisson. Who thinks best depends on how long you let them: Budget-dependent rankings in LLM evaluation, 2026. arXiv preprint arXiv:2608.12150.

Yifan Dou, Shikan Lian, and Shibo Li. Conformal cascade: Distribution-free accuracy guarantees for multi-tier llm inference, 2026. arXiv preprint arXiv:2607.25018v3.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Chenrui Fan, Yize Cheng, Ming Li, Yongyuan Liang, Tianyi Zhou, and Soheil Feizi. Thinking hard, not smart: Reasoning models fail to ration test-time compute across questions, 2026. arXiv preprint arXiv:2608.07968v2.

Dylan J. Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3199–3210, 2020. arXiv:2002.04926.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, 2016. arXiv:1602.04589.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025. doi: 10.18653/v1/2025.findings-acl.1274.

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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

Jenny Y. Huang, Mehul Damani, Yousef El-Kurdi, Ramon Astudillo, and Wei Sun. Latency and token-aware test-time compute, 2025. arXiv preprint arXiv:2509.09864v1.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, 2014. arXiv:1312.7308.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. The multi-fidelity multi-armed bandit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1610.09726.

Mobina Kashaniyan and Ali Jannesari. Interpretable adaptive sampling for llm test-time scaling, 2026. arXiv preprint arXiv:2608.03961v1.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. arXiv:1407.4443.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Jungseob Lee, Seongtae Hong, Seungjun Lee, Jaehyung Seo, Junyoung Son, Sugyeong Eo, Chanjun Park, Hyeongju Park, Hyeonseok Moon, and Heuiseok Lim. Dart: Draft-agreement routing for training-free adaptive thinking budgets in hybrid reasoning models, 2026. arXiv preprint arXiv:2606.23181v1.

Junyan Li, Wenshuo Zhao, Yang Zhang, and Chuang Gan. Steering llm thinking with budget guidance, 2025. arXiv preprint arXiv:2506.13752.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 661–670, 2010. doi: 10.1145/1772690.1772758.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. arXiv:1603.06560.

Yifei Li, Zihui Gao, and Laks V. S. Lakshmanan. Wiserouter: Llm routing with workload budget constraint, 2026. arXiv preprint arXiv:2607.23765v1.

Baohao Liao, Hanze Dong, Yuhui Xu, Doyen Sahoo, Christof Monz, Junnan Li, and Caiming Xiong. Fractured chain-of-thought reasoning, 2025. arXiv preprint arXiv:2505.12992.

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. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Xiaodong Lu, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Zhijun Chen, Yu Luo, Fuzhen Zhuang, Yikun Ban, and Deqing Wang. Contextual rollout bandits for reinforcement learning with verifiable rewards, 2026. arXiv preprint arXiv:2602.08499v2.

Thodoris Lykouris, Eva Tardos, and Drishti Wali. Feedback graph regret bounds for thompson sampling and ucb, 2019. ALT 2020; arXiv:1905.09898.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. arXiv:1106.2436.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data, 2024. arXiv:2406.18665.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Optimizing anytime reasoning via budget relative policy optimization, 2025. arXiv preprint arXiv:2505.13438.

Ritik Raj, Souvik Kundu, Sarbartha Banerjee, Dheemanth Joshi, Ishita Vohra, and Tushar Krishna. Trace-router: Task-consistent and adaptive online routing for agentic ai, 2026. arXiv preprint arXiv:2607.22465v2.

Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Indexed minimum empirical divergence for unimodal bandits, 2021. NeurIPS 2021; arXiv:2112.01452.

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.

Ajay Narayanan Sridhar, Ronak Singh, Mehrdad Mahdavi, and Vijaykrishnan Narayanan. Correlation-aware contextual bandits with surrogate rewards for llm routing, 2026. arXiv preprint arXiv:2607.09015v1.

Chloe H. Su, Zhe Ye, Samuel Tenka, Aidan Yang, Soonho Kong, and Udaya Ghai. Learning adaptive llm decoding, 2026. arXiv preprint arXiv:2603.09065v2.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, and Hanjie Chen. Stop overthinking: A survey on efficient reasoning for large language models, 2025. arXiv:2503.16419.

Annette Taberner-Miller. Paretobandit: Budget-paced adaptive routing for non-stationary llm serving, 2026. arXiv preprint arXiv:2604.00136v2.

Chanakya Varude, Jay Chaudhary, Siddharth Kaushik, and Prasanna Chaporkar. Threshold-based optimal arm selection in monotonic bandits: Regret lower bounds and algorithms, 2025. arXiv preprint arXiv:2509.02119v1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc 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 (ICLR)*, 2023. arXiv:2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. arXiv:2406.01574.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, and Zhuosheng Zhang. Thoughts are all over the place: On the underthinking of o1-like llms, 2025. arXiv:2501.18585.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.11903.

Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A. Rossi, and Hoda Eldardiry. Learning to route llms from bandit feedback: One policy, many trade-offs, 2025. arXiv preprint arXiv:2510.07429v1.

Yuxiao Wen, Yanjun Han, and Zhengyuan Zhou. Stochastic contextual bandits with graph feedback: From independence number to mas number, 2024. arXiv preprint arXiv:2402.18591.

Huasen Wu, R. Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits, 2015. NeurIPS 2015; arXiv:1504.06937.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.10601.

Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 41–48, 2011.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.03094.

Jiahao Zeng, Ming Tang, and Ningning Ding. Learning to route llms from implicit cost-performance preferences via meta-learning, 2026. arXiv preprint arXiv:2606.06178v1.

Zhiyuan Zhai, Bingcong Li, Bingnan Xiao, Ming Li, and Xin Wang. Adaptive test-time compute allocation for reasoning llms via constrained policy optimization, 2026. arXiv preprint arXiv:2604.14853v1.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach, 2025. arXiv preprint arXiv:2506.12721v2.

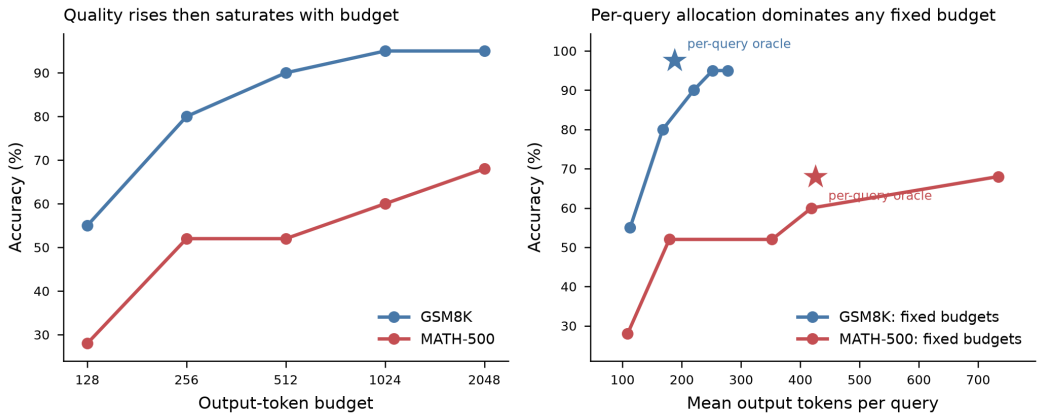


Figure 2: The allocation opportunity in our own measurements ($n = 40$ GSM8K, $n = 25$ MATH-500, self-report grading). **Left:** quality rises then saturates with the output-token budget, and the saturation point differs by domain. **Right:** the per-query cheapest-correct oracle ($\star$) lies above and to the left of the fixed-budget curve — higher accuracy at fewer mean tokens. The oracle requires per-problem correctness in hindsight and is not achievable; it bounds what any allocator could win. §5 reports how much of this gap learned policies actually capture, which on these two domains is none.

## A THE ALLOCATION OPPORTUNITY

**The oracle-versus-fixed gap depends on the grading instrument.** Measured against the cheapest highest-accuracy fixed rung, the cheapest-correct oracle on the *self-report* path saves $25.4\%$ of tokens while gaining $2.5$ accuracy points on GSM8K ($187.8$ vs. $251.9$ tokens, $0.975$ vs. $0.950$) and saves $42.0\%$ at equal accuracy on MATH-500 ($425.8$ vs. $733.8$, both $0.680$); it dominates on both. On the *forced-answer* path the same comparison gives a $30.2\%$ saving at $+8.6$ accuracy points on MATH-500 and $18.2\%$ at equal accuracy on MBPP (execution-graded), but $-2.3\%$ on GSM8K at $+10$ points, because that path's best rung is already $b$=256, the cheapest useful one — the oracle still gains accuracy there but no longer saves tokens. Both readings are reported because the direction of the token term is instrument-dependent while the accuracy term is not, and §5 scores policies on each universe separately for the same reason.

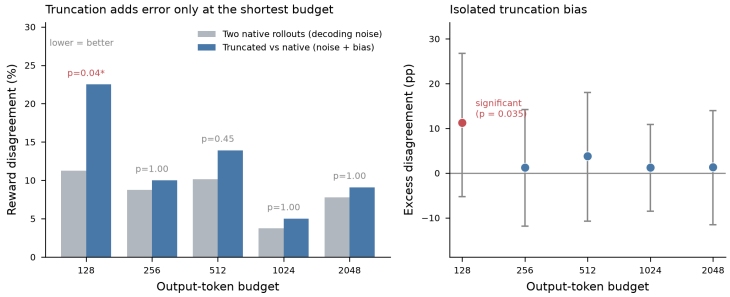


Figure 3: The fidelity measurement. **Left:** per-budget reward disagreement for two native rollouts (grey, the decoding-noise floor) and for a truncated versus a native rollout (blue, noise plus bias), with exact-binomial McNemar $p$-values on the paired per-problem agreement. **Right:** the isolated excess, with 95% intervals. Only $b = 128$ separates from its control; the interval on its magnitude nonetheless spans zero, so the figure establishes the direction of that bias and not its size.

## B MEASUREMENT COVERAGE: WHAT IS MEASURED AND WHAT IS BLANK

Every unmeasured cell below is blank, never zero. The binding constraint on scale was API wall-clock latency (roughly 45–135 seconds per generation) and then a per-session token ceiling, not tokens per call.

Table 3: Coverage by domain. 400 generation arms per domain (40 problems $\times$ 5 budgets $\times$ 2 replicates). The two count columns are distinct and add up: *extractions* counts arms that received a live forced-answer extraction call, *empty-gen* counts arms that hit the cap without emitting anything (graded 0 — a genuine "no answer at this budget" observation, requiring no extraction call), and their sum is the number of graded forced-answer rows, which is what the per-budget $n$ of Table 5 reports. GSM8K is complete ($396 + 4 = 400$); MATH-500 is partial ($278 + 77 = 355$ of 400, the remaining 45 arms lost to the session token ceiling), which costs 5 problems their complete two-replicate ladder and is why PRIMARY has $n = 75$ rather than 80. "Forced-answer accuracy" is UNMEASURED for MBPP and MMLU-Pro: live traces exist but received zero forced-answer extraction calls, so those 20 summary rows carry a blank accuracy and are excluded by domain from PRIMARY. Their raw traces *are* graded on the execution / self-report path, which is where their numbers in §5 come from.

| domain | problems (per rep.) | gen. arms | extrac-tions | empty-gen | graded rows | noise floor measured | forced-answer accuracy |
|---|---|---|---|---|---|---|---|
| GSM8K | 40 | 400 | 396 | 4 | 400 | forced-answer + self/exec | measured |
| MATH-500 | 40 | 400 | 278 | 77 | 355 | forced-answer + self/exec | measured (partial) |
| MBPP | 40 | 400 | 0 | – | 0 | self/exec path only | **UNMEASURED** |
| MMLU-Pro | 40 | 400 | 0 | – | 0 | self/exec path only | **UNMEASURED** |

Decision-time triage features are present for 100% of GSM8K problems, 50% of MATH-500, and 0% of MBPP and MMLU-Pro. The BREADTH universe therefore runs feature-based policies on free features only, enforced by an assertion that fails if any missing value reaches a design matrix rather than by convention.

## C THE FOUR-DOMAIN DECODING-NOISE FLOOR

Table 4 gives replicate agreement for all four domains on a path that costs no additional calls: the self-reported answer line for math and multiple-choice, and execution of the dataset's own unit tests for MBPP. Rewards are unconditional — no answer inside the cap scores zero. **The noise floor is substantial everywhere**: 0–25% replicate disagreement. This is the context in which any claim that a surrogate "agrees 90% of the time" has to be read.

## D ACCURACY AND COST OF EVERY LADDER RUNG

Table 5 is the table the replay of §5 runs on. Non-monotonicity in budget is measurable directly from it: $7/40$ GSM8K and $3/35$ MATH-500 problems flip from correct to wrong as the budget grows (17.5% and 8.6%), which is why no result in this paper assumes reward is monotone in compute.

**Cheapest-correct oracle versus the best fixed budget.** Reading directly off Table 5's underlying per-problem rows: MATH-500 shows a 30.2% token saving at $+8.6$ accuracy points on the forced-answer path, and MBPP an 18.2% saving at equal accuracy under execution grading. GSM8K shows *no* saving ($-2.3\%$) because its best fixed budget is already the cheapest useful rung, $b = 256$ — which bounds where per-query allocation can help and is reproduced as no learnable gain in §5.

**Decision-time features and their predictive validity.** All context features are computable before generation, so none spends an arm. Deterministic and dataset-provided features (problem length, option count, test count, dataset level and category) are free. Three optional triage features come from one cheap side-call per problem that is instructed not to solve the problem; that call is a fixed side-cost independent of which arm is pulled, but it is *not* free and is charged in all cost accounting.

Table 4: Independent-versus-independent replicate agreement, $n = 40$ problems per domain and budget, Wilson intervals. The two negative $\kappa$ values on MMLU-Pro at $b \in \{128, 256\}$ arise because accuracy there is near zero and near-constant (both replicates almost always wrong), which makes chance agreement essentially total; the raw agreement is the interpretable quantity in those two cells.

| domain | path | $b$ | agreement | 95% CI | disagreement | $\kappa$ |
|---|---|---|---|---|---|---|
| GSM8K | self-report | 128 | 0.775 | $[0.625, 0.877]$ | 0.225 | 0.486 |
| GSM8K | self-report | 256 | 0.950 | $[0.835, 0.986]$ | 0.050 | 0.875 |
| GSM8K | self-report | 512 | 0.925 | $[0.801, 0.974]$ | 0.075 | 0.776 |
| GSM8K | self-report | 1024 | 0.925 | $[0.801, 0.974]$ | 0.075 | 0.793 |
| GSM8K | self-report | 2048 | 0.900 | $[0.770, 0.960]$ | 0.100 | 0.688 |
| MATH-500 | self-report | 128 | 0.975 | $[0.871, 0.996]$ | 0.025 | 0.942 |
| MATH-500 | self-report | 256 | 0.775 | $[0.625, 0.877]$ | 0.225 | 0.548 |
| MATH-500 | self-report | 512 | 0.775 | $[0.625, 0.877]$ | 0.225 | 0.557 |
| MATH-500 | self-report | 1024 | 0.900 | $[0.770, 0.960]$ | 0.100 | 0.787 |
| MATH-500 | self-report | 2048 | 0.950 | $[0.835, 0.986]$ | 0.050 | 0.890 |
| MBPP | unit tests | 128 | 0.825 | $[0.681, 0.913]$ | 0.175 | 0.622 |
| MBPP | unit tests | 256 | 0.900 | $[0.770, 0.960]$ | 0.100 | 0.750 |
| MBPP | unit tests | 512 | 0.925 | $[0.801, 0.974]$ | 0.075 | 0.727 |
| MBPP | unit tests | 1024 | 0.925 | $[0.801, 0.974]$ | 0.075 | 0.625 |
| MBPP | unit tests | 2048 | 1.000 | $[0.912, 1.000]$ | 0.000 | 1.000 |
| MMLU-Pro | self-report | 128 | 0.950 | $[0.835, 0.986]$ | 0.050 | $-0.026$ |
| MMLU-Pro | self-report | 256 | 0.925 | $[0.801, 0.974]$ | 0.075 | $-0.035$ |
| MMLU-Pro | self-report | 512 | 0.925 | $[0.801, 0.974]$ | 0.075 | 0.625 |
| MMLU-Pro | self-report | 1024 | 0.750 | $[0.598, 0.858]$ | 0.250 | 0.494 |
| MMLU-Pro | self-report | 2048 | 0.900 | $[0.770, 0.960]$ | 0.100 | 0.713 |

Their signal is real but weak: Spearman $|r|$ against reward is 0.16–0.26 ($p < 0.03$) on GSM8K and MATH-500, and predicted tokens correlate with actual tokens at $r = 0.31$ (GSM8K) and 0.44 (MATH-500). Problem length alone is useless ($p > 0.29$), so a length-only router has nothing to condition on. A short-generation confidence signal would violate the no-arm-spending constraint — obtaining it means pulling the $b = 128$ arm — and was deliberately excluded.

## E Replay policies and references

**Policies.** (1) *Fixed budget* at each rung. (2) *Confidence threshold escalation*: start at 128 and climb while a calibrated confidence score is below threshold, paying the cumulative cost of every rung pulled; the score is $P(\text{correct} \mid \text{post-pull observables})$ fit by logistic regression on a disjoint problem set using only what a serving stack sees for free — whether an answer was extractable, whether the model terminated on its own rather than hitting the cap, and how much cap headroom remained. Threshold swept over 7 values. (3) *Supervised router*, two variants trained on a disjoint 50% problem split: one predicts the cheapest-correct rung directly, one predicts per-rung correctness probability and picks the utility-maximizing rung. Held-out numbers are the ones reported. (4) *Contextual bandit, bandit feedback only*: LinUCB and LinTS over the decision-time features, each pull updating only the arm pulled. (5) *Same + truncation feedback*: each pull of arm $k$ additionally updates every level $j < k$ from the measured truncated observation. (6) *Same + trust threshold*: reuse truncated observations only for levels $\geq 256$, motivated by §4.

**References.** Per-query oracle ($\lambda$-optimal rung per problem, an upper bound); cheapest-correct oracle; random rung; and the **best fixed rung per domain chosen in hindsight**, which was added during analysis because it turned out to be the binding reference — any policy that beats it is doing genuine per-query allocation rather than domain identification.

**Tuning discipline.** Exploration rate for LinUCB/LinTS was tuned on a disjoint set of tuning seeds (200–211) and then frozen at a single shared value for every reported bandit, so no learner was tuned on its own evaluation seeds. Policies that read a triage feature are charged its 75 output tokens per problem, priced as generator output tokens — deliberately conservative against feature-based

Table 5: Accuracy and mean measured completion tokens by grading path, domain and budget, independent arms only, interpretable rows only. Note GSM8K saturates by $b = 256$ on the forced-answer path, which is why per-query allocation has almost nothing to win there, and note MMLU-Pro's climb from $2.5\%$ to $77.5\%$, which is partly a failure-to-terminate effect rather than a reasoning effect (§7).

| path | domain | $b$ | $n$ | accuracy | 95% CI | mean tokens |
|---|---|---|---|---|---|---|
| unit tests | MBPP | 128 | 80 | 0.638 | $[0.528, 0.734]$ | 96.4 |
| unit tests | MBPP | 256 | 80 | 0.725 | $[0.619, 0.811]$ | 143.9 |
| unit tests | MBPP | 512 | 80 | 0.838 | $[0.742, 0.903]$ | 197.5 |
| unit tests | MBPP | 1024 | 80 | 0.888 | $[0.800, 0.940]$ | 304.0 |
| unit tests | MBPP | 2048 | 80 | 0.925 | $[0.846, 0.965]$ | 352.1 |
| forced-answer | GSM8K | 128 | 80 | 0.725 | $[0.619, 0.811]$ | 119.9 |
| forced-answer | GSM8K | 256 | 80 | 0.838 | $[0.742, 0.903]$ | 175.9 |
| forced-answer | GSM8K | 512 | 80 | 0.838 | $[0.742, 0.903]$ | 235.2 |
| forced-answer | GSM8K | 1024 | 80 | 0.800 | $[0.700, 0.873]$ | 268.7 |
| forced-answer | GSM8K | 2048 | 80 | 0.838 | $[0.742, 0.903]$ | 280.3 |
| forced-answer | MATH-500 | 128 | 75 | 0.333 | $[0.237, 0.446]$ | 115.9 |
| forced-answer | MATH-500 | 256 | 70 | 0.543 | $[0.427, 0.654]$ | 197.3 |
| forced-answer | MATH-500 | 512 | 70 | 0.529 | $[0.413, 0.641]$ | 333.6 |
| forced-answer | MATH-500 | 1024 | 70 | 0.614 | $[0.497, 0.720]$ | 475.3 |
| forced-answer | MATH-500 | 2048 | 70 | 0.643 | $[0.526, 0.745]$ | 634.2 |
| self-report | GSM8K | 128 | 80 | 0.313 | $[0.222, 0.421]$ | 119.9 |
| self-report | GSM8K | 256 | 80 | 0.725 | $[0.619, 0.811]$ | 175.9 |
| self-report | GSM8K | 512 | 80 | 0.788 | $[0.686, 0.863]$ | 235.2 |
| self-report | GSM8K | 1024 | 80 | 0.763 | $[0.659, 0.842]$ | 268.7 |
| self-report | GSM8K | 2048 | 80 | 0.800 | $[0.700, 0.873]$ | 280.3 |
| self-report | MATH-500 | 128 | 80 | 0.313 | $[0.222, 0.421]$ | 111.2 |
| self-report | MATH-500 | 256 | 80 | 0.463 | $[0.358, 0.571]$ | 188.9 |
| self-report | MATH-500 | 512 | 80 | 0.488 | $[0.381, 0.595]$ | 323.8 |
| self-report | MATH-500 | 1024 | 80 | 0.625 | $[0.516, 0.723]$ | 451.7 |
| self-report | MATH-500 | 2048 | 80 | 0.650 | $[0.541, 0.746]$ | 587.9 |
| self-report | MMLU-Pro | 128 | 80 | 0.025 | $[0.007, 0.087]$ | 127.5 |
| self-report | MMLU-Pro | 256 | 80 | 0.038 | $[0.013, 0.105]$ | 254.4 |
| self-report | MMLU-Pro | 512 | 80 | 0.113 | $[0.060, 0.200]$ | 506.7 |
| self-report | MMLU-Pro | 1024 | 80 | 0.425 | $[0.323, 0.534]$ | 892.5 |
| self-report | MMLU-Pro | 2048 | 80 | 0.775 | $[0.672, 0.853]$ | 1229.5 |

policies, since triage runs on a much cheaper model. The $0.2\times$ row of the triage-price ablation is the more realistic price and changes no conclusion's sign.

## F Q3 HEADROOM DETAIL, FRONTIERS, AND ABLATIONS

**Detail dropped from §5 for space.** Headroom fractions are a *ratio of means* — mean paired advantage over the best fixed budget divided by the mean oracle gap — rather than a mean of per-seed ratios, because a per-seed gap can be zero; 148 non-finite per-seed ratios in an earlier version of the estimator are what prompted the change, which shifted every headroom number by $0.01$–$0.05$ and reversed no conclusion. Confidence escalation reaches the matched-quality target on several domains but always at a large negative saving, because it pays for every rung it climbs; that is the honest accounting for sequential escalation and means its frontier position reflects cumulative cost rather than poor rung selection. Quality-at-matched-spend gains not quoted in the main text: $+0.152$ on BREADTH GSM8K, $+0.107$ on MBPP, $+0.091$ on MATH-500. Cross-domain transfer, router utility at $\lambda = 0.5$: GSM8K $0.743$ in-domain against $0.656$ trained on MATH-500; MATH-500 $0.437$ in-domain against $0.369$ trained on GSM8K, i.e. from $-0.05/-0.12$ of within-test-domain headroom in-domain to $-0.51/-0.97$ cross-domain. Problem length alone carries essentially no signal ($|\text{Spearman } r| < 0.11$, $p > 0.14$), so a length-only router would have nothing to condition on.

Table 6: Full Q1/Q2 comparison: paired per-seed utility deltas with 95% CIs over 50 seeds and Holm-corrected $p$-values within each seven-$\lambda$ family. This is the complete version of Table 2.

| | LinUCB | | | LinTS | | |
|---|---|---|---|---|---|---|
| $\lambda$ | $\Delta$ | 95% CI | Holm $p$ | $\Delta$ | 95% CI | Holm $p$ |
| *Q1: bandit + truncation reuse − bandit feedback only* | | | | | | |
| 0 | $-0.0033$ | $[-0.0170, +0.0105]$ | 0.79 | $-0.0076$ | $[-0.0156, +0.0004]$ | 0.27 |
| 0.1 | $-0.0048$ | $[-0.0136, +0.0040]$ | 0.79 | $-0.0045$ | $[-0.0137, +0.0046]$ | 0.67 |
| 0.25 | $-0.0054$ | $[-0.0131, +0.0023]$ | 0.71 | $-0.0002$ | $[-0.0083, +0.0080]$ | 0.97 |
| 0.5 | $+0.0040$ | $[-0.0029, +0.0110]$ | 0.79 | $+0.0046$ | $[-0.0021, +0.0114]$ | 0.56 |
| 1.0 | $+0.0189$ | $[+0.0105, +0.0272]$ | 0.0003 | $+0.0160$ | $[+0.0074, +0.0245]$ | 0.0031 |
| 2.0 | $+0.0341$ | $[+0.0238, +0.0444]$ | $< 10^{-6}$ | $+0.0426$ | $[+0.0311, +0.0541]$ | $< 10^{-7}$ |
| 4.0 | $+0.0384$ | $[+0.0281, +0.0486]$ | $< 10^{-7}$ | $+0.0420$ | $[+0.0299, +0.0541]$ | $< 10^{-7}$ |
| *Q2: trust threshold − naive reuse of all levels* | | | | | | |
| 0 | $+0.0070$ | $[-0.0033, +0.0173]$ | 0.19 | $+0.0085$ | $[-0.0013, +0.0183]$ | 0.19 |
| 0.1 | $+0.0082$ | $[-0.0002, +0.0167]$ | 0.12 | $+0.0126$ | $[+0.0030, +0.0223]$ | 0.039 |
| 0.25 | $+0.0130$ | $[+0.0050, +0.0211]$ | 0.013 | $+0.0111$ | $[+0.0039, +0.0182]$ | 0.019 |
| 0.5 | $+0.0117$ | $[+0.0062, +0.0171]$ | 0.0007 | $+0.0128$ | $[+0.0068, +0.0188]$ | 0.0007 |
| 1.0 | $+0.0076$ | $[+0.0020, +0.0133]$ | 0.042 | $+0.0095$ | $[+0.0034, +0.0156]$ | 0.019 |
| 2.0 | $-0.0089$ | $[-0.0159, -0.0019]$ | 0.047 | $-0.0048$ | $[-0.0119, +0.0023]$ | 0.19 |
| 4.0 | $-0.0185$ | $[-0.0262, -0.0109]$ | 0.0001 | $-0.0184$ | $[-0.0266, -0.0103]$ | 0.0004 |

Table 7: Fraction of the (oracle − best fixed budget) utility gap captured, held-out scoring, all domains pooled. Estimated as a ratio of means — mean paired advantage over the best fixed budget divided by the mean oracle gap — not as a mean of per-seed ratios, because a per-seed gap can be zero; where the mean gap itself falls below 0.02 utility units the fraction is blanked rather than reported as a large ratio, and the paired numerator is always interpretable. Negative fractions mean the policy is worse than the best fixed budget.

| policy | universe | $\lambda$=0.25 | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|---|
| best fixed rung *per domain* (hindsight) | BREADTH | 0.026 | 0.067 | 0.228 | 0.318 |
| supervised router | BREADTH | $-0.046$ | **0.308** | **0.445** | **0.525** |
| bandit, pulled arm only | BREADTH | $-0.882$ | $-0.326$ | 0.035 | 0.014 |
| best fixed rung per domain | PRIMARY | 0.032 | 0.000 | 0.000 | 0.000 |
| supervised router | PRIMARY | $-0.139$ | $-0.478$ | $-0.916$ | $-1.115$ |
| bandit + trusted truncation feedback | PRIMARY | $-0.095$ | $-0.419$ | $-0.869$ | $-1.159$ |

## G EXTENDED RELATED WORK: THREE PRECISE DISTINCTIONS

**Multi-fidelity bandits (Kandasamy et al., 2016).** An arm may be played at one of $M$ ordered fidelities; fidelity $m$ costs less than the top fidelity and returns a *biased* estimate of it, and MF-UCB beats fidelity-blind strategies with a matching lower bound. This is the same ordered-cost, biased-cheap-observation structure we face, and it is the prior art any cost-aware claim on a budget ladder must be measured against. The one surviving structural difference is **nesting**: there each fidelity is a separately paid pull, whereas here one paid rollout at level $k$ yields every level $j < k$ simultaneously. Proposition 6 prices that difference exactly — $\rho_K = \sum_k b_k / b_K$, bounded by $r/(r-1)$ for a geometric ladder and equal to 1.94 on ours — and it is a bounded constant, never a rate improvement. A further difference matters for the analysis: their fidelity gaps are *known*, and their lowest fidelity is still an unbiased draw of its own distribution, whereas our per-level bias has unknown sign and bounds a quantity that Proposition 13 shows is not identifiable from truncated observations.

**Correlation-aware contextual bandits for LLM routing (Sridhar et al., 2026).** This is the closest contemporary work on the feedback axis: a contextual bandit over heterogeneous models in which an external affinity oracle supplies a time-varying directed feedback graph with noisy surrogate rewards for unpulled arms, with regret bounded through the graph's independence number and a specialization to transitive graphs. The setting differs from ours in where the graph comes from

Table 8: Feature ablation: which block carries the signal. Utility of the utility-maximizing supervised router, held-out scoring, PRIMARY, by feature set. **Bold marks the best feature set in each column.** The three triage features cost 75 output tokens per problem, charged at parity with generator tokens; the free features (problem length, dataset difficulty level) and the domain indicator cost nothing. Triage helps only at $\lambda = 0.25$ and is progressively dominated as tokens are priced, which is a statement about its price rather than its signal — the measurement track's own validity numbers for it are |Spearman $r$| = 0.16–0.26. The domain indicator alone is best at $\lambda = 2.0$, where the cheap rung is optimal for both domains and knowing which domain one is in is nearly sufficient.

| feature set | $\lambda$=0.25 | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|
| free only (problem tokens, dataset level) | 0.6429 | **0.5853** | **0.4919** | 0.3366 |
| free + domain indicator | 0.6442 | 0.5847 | 0.4907 | 0.3350 |
| domain indicator only | 0.6426 | 0.5836 | 0.4847 | **0.3453** |
| free + triage (no domain) | **0.6696** | 0.5783 | 0.4279 | 0.1930 |
| triage + domain indicator | 0.6675 | 0.5791 | 0.4330 | 0.2022 |
| free + triage + domain (all) | 0.6667 | 0.5759 | 0.4291 | 0.1919 |

Table 9: Token saving at matched top quality, held-out. The target is the quality of the highest-quality fixed budget in that cell. **No learned policy achieves a saving on any cell.** The oracle's 11–64% quantifies the headroom that exists in the measured table; fitted methods do not convert it into savings at the top quality point. Confidence escalation reaches the target on several cells but always at a large negative saving, because it pays for every rung it climbs.

| universe | domain | target | oracle saving | best learned policy |
|---|---|---|---|---|
| PRIMARY | GSM8K | 0.852 @ 174 tok ($b$=256) | +11.1% | router hits target at 287 tok (−64.7%) |
| PRIMARY | MATH-500 | 0.652 @ 641 tok ($b$=2048) | +64.1% | none reaches the target quality |
| BREADTH | GSM8K | 0.817 @ 275 tok | +35.2% | none reaches the target quality |
| BREADTH | MATH-500 | 0.662 @ 588 tok | +61.4% | none reaches the target quality |
| BREADTH | MBPP | 0.923 @ 361 tok | +52.8% | none reaches the target quality |
| BREADTH | MMLU-Pro | 0.789 @ 1216 tok | +22.8% | router matches at 1216 tok (0.0%) |

rather than in how it is used — their arms are distinct models with no ordering and an externally supplied graph, ours are nested compute budgets on one model whose graph is intrinsic to the arm structure. Their surrogate bias is a constant per arm and hence removable by de-biasing; the bias we measure is bidirectional across problems (§4), so there is no constant offset to remove, and their decoupled estimator is not a drop-in fix.

**Semi-bandit learning for monotone stochastic optimization (Agarwal et al., 2023).** This work's abstract states an extension to censored and binary feedback in which the policy observes truncated or thresholded versions of the probed variables, which reads as though it subsumes our setting; we flag it explicitly because a reviewer may have read only the abstract. Reading the full text resolves it. Its monotonicity (Definition 1.1) is monotonicity of the *offline optimum* under stochastic dominance of $n$ *independent* item distributions, not reward monotone in an ordered arm index; its semi-bandit feedback means the realizations of whichever items the policy probed, as in series testing that halts at the first failure; and the words "censor", "truncat" and "context" each occur zero times in the body. Our $K$ levels are *dependent views of one latent trace* and our setting is contextual, which their framework has no notion of. We therefore cite it as the nearest monotone-semi-bandit relative and state the distinction rather than claiming a result against it.

**Adaptive inference in serving systems.** Most LLM applications of bandit machinery treat model identity as the arm — routing and cascades trading quality against price, with online variants adding workload budget constraints, non-stationary pricing, agentic routing, preference-conditioned policies, and distribution-free cascade guarantees (Chen et al., 2023; Yue et al., 2024; Ong et al., 2024; Chen et al., 2024; Li et al., 2026; Taberner-Miller, 2026; Raj et al., 2026; Wei et al., 2025; Zeng et al., 2026; Dou et al., 2026) — while bandit treatments of rollout scheduling optimise which training rollouts to learn from rather than how much inference compute to spend (Lu et al., 2026; Kashaniyan & Jannesari, 2026; Su et al., 2026). Quality scaling by repeated sampling and verification (Brown

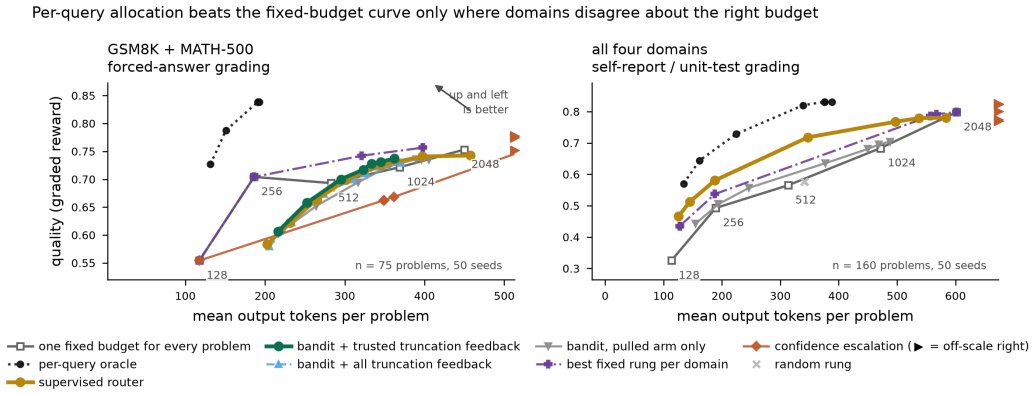


Figure 4: Quality–cost frontiers, held-out scoring. Per-query allocation beats the fixed-budget curve only in the four-domain universe, where domains disagree about the right rung; in the two-math universe (left) every learned policy sits on or below the fixed-budget curve while the per-query oracle (dotted) is far above all of them.

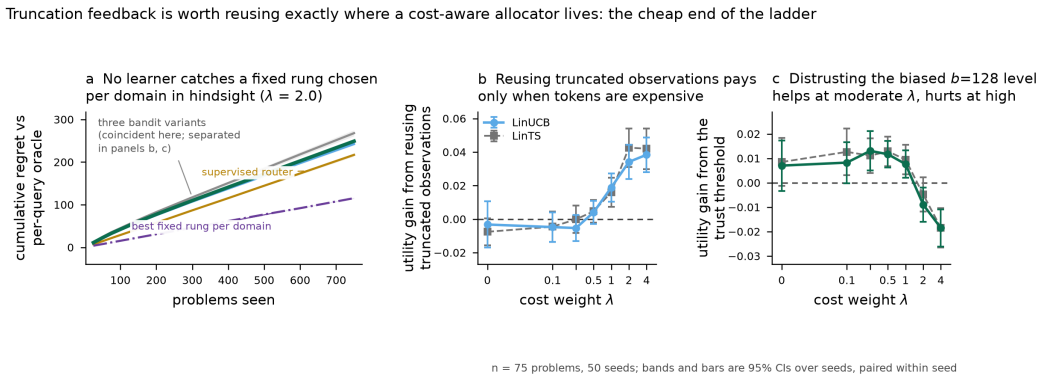


Figure 5: (a) No learner catches a hindsight-chosen per-domain fixed rung at $\lambda = 2.0$. (b) Reusing truncated observations pays only when tokens are expensive; every $\lambda \leq 0.5$ interval spans zero. (c) The trust threshold helps at moderate $\lambda$ and *hurts* at high $\lambda$, where the distrusted rung is itself optimal. Bars and bands are 95% CIs over 50 seeds, paired within seed; these control replay-stream variance, not problem-sampling variance. Figure 4 in the appendix gives the corresponding quality–cost frontiers.

et al., 2024; Cobbe et al., 2021; Lightman et al., 2024; Wang et al., 2023; Wei et al., 2022; Yao et al., 2023) supplies the reward model our ladder scores against, and de Souza & Panisson (2026) report budget-dependent rank reversals consistent with the per-instance non-monotonicity we measure.

**Pre-LLM ancestry.** A reader from the bandit community may reasonably read our setup as a budget ladder with early stopping in the style of Hyperband (Li et al., 2018), which casts exactly that as a pure-exploration bandit. The difference is that Hyperband's ladder allocates across *configurations* under a fixed-confidence objective with no context and no per-round cost-weighted regret, and its rungs are separately paid. We regard the lineage as real and worth stating rather than worth arguing with.

**Further context.** Pure-exploration budget ladders, cost-constrained bandits, and the standard contextual toolkit supply the classical background (Li et al., 2018; Even-Dar et al., 2006; Kaufmann et al., 2016; Jamieson et al., 2014; Garivier & Kaufmann, 2016; Badanidiyuru et al., 2013; Wu et al., 2015; Agrawal & Devanur, 2016; Langford & Zhang, 2007; Li et al., 2010; Chu et al., 2011; Agrawal & Goyal, 2013; Abbasi-Yadkori et al., 2011; Foster & Rakhlin, 2020; Lattimore & Szepesvári,

2020). Bandits for LLM inference that use *model identity* rather than budget level as the arm, and the quality-scaling methods our ladder scores against, supply the rest (Chen et al., 2023; Yue et al., 2024; Ong et al., 2024; Chen et al., 2024; Li et al., 2026; Taberner-Miller, 2026; Raj et al., 2026; Wei et al., 2025; Zeng et al., 2026; Dou et al., 2026; Lu et al., 2026; Kashaniyan & Jannesari, 2026; Su et al., 2026; Brown et al., 2024; Cobbe et al., 2021; Lightman et al., 2024; Wang et al., 2023; Wei et al., 2022; Yao et al., 2023; de Souza & Panisson, 2026).

# H LIMITATIONS, EXPANDED

**Grader and metric artifacts.** The exact-match grader under-credits algebraically equivalent MATH-500 forms, so MATH-500 accuracy is a lower bound on mathematical correctness. The failure-to-terminate effect that inflates MMLU-Pro's budget sensitivity also affects MATH-500 at $b = 128$, where only $36\%$ of rollouts emitted an answer inside the cap; rewards are unconditional throughout, so a rollout that reasons correctly but does not commit to an answer scores zero. That is the right convention for a serving system, which cannot use an answer it did not receive, but it means part of what our ladder measures is answer-emission behaviour rather than reasoning quality.

**Why the seed intervals are narrower than the truth.** Each replay stream resamples problems *with replacement from the same fixed set* of 75 or 160 problems, so the 50-seed intervals quantify variation in the order problems arrive and in which measured replicate is drawn — not variation in which problems the benchmark contains. A study with the same design on a fresh problem sample could differ by considerably more than our intervals suggest. We report them because they are the correct uncertainty for the replay comparison, not because they bound generalization.

**What a second model would settle.** The single-model scope is the most consequential limitation, because the mechanism we propose for the $b = 128$ bias — that the model plans against the budget it is given — predicts that the bias should depend on how strongly a model conditions on its budget. A model trained with explicit budget guidance should show a *larger* truncation bias, and a model that ignores its budget should show none. Testing that requires no new measurement design, only a second generator.

**Theory gaps in full.** Beyond the summary in §7: realizability of a per-level linear quality model is assumed, and $n = 65$–$80$ can neither fit nor falsify it, though standard misspecification of size $\eta$ adds only $O(\eta\sqrt{d}\,T)$. Context diversity $\lambda_{\min}(\Sigma) \geq \sigma^2 > 0$ is a property of the query stream and is used *only* by the pooled branch of Theorem 8, so the bound survives its failure through the native branch. The lower bounds we do have — $\Omega(\sqrt{T})$ for $K = 2$ with full information, $\Omega(\delta_j T)$ for truncation-only algorithms (Proposition 13), $\Omega(\sqrt{dT})$ for the native channel — bracket but do not match the adaptive minimum; an algorithm that estimates $\delta_k(x)$ from sparse native pulls might interpolate better, and we state that as an open problem rather than asserting optimality. Proposition 14's three-step mechanism is proved but its linear-growth rate is measured (doubling ratio 3.33) rather than derived in closed form, and under a single *uniform* bias the capped rule is indistinguishable from the default — the failure needs the measured per-level profile. Lemma 18 bounds $|\langle \phi_t, \xi_k \rangle|$ but gives no sign, because $\sum_s (\phi_t^\top V^{-1} \phi_s)\delta_s$ can be negative even when every $\delta_s \geq 0$. The explicit factor 2 in Theorem 8 absorbs burn-in and union-bound constants that are not optimized. Finally, the theory's simulation is calibrated to our measurement and therefore inherits its single-model scope.

# I PROBLEM FORMULATION

## I.1 COMPUTE LADDERS AS A NESTED-OBSERVATION BANDIT

**Ordered budget levels.** Fix a single language model $\pi$ and an increasing ladder of decoding budgets $b_1 < b_2 < \cdots < b_K$, measured in generated tokens; in our experiments $\mathcal{B} = \{128, 256, 512, 1024, 2048\}$. The ordered discrete budget set on a fixed backbone is not new: it is exactly the action space of Zhai et al. (2026), which orders $\mathcal{B}$ by cost and solves the allocation problem offline. Our departure is that the level is chosen *online, per query*, and that the learner exploits the nesting described below.

**Interaction protocol.** At round $t = 1, \dots, T$ the learner receives a query with context $x_t \in \mathcal{X}$, drawn i.i.d. from a fixed distribution $\mathcal{P}_x$. The context is whatever is computable *before* generation — prompt length, embedding features, a difficulty probe in the sense of Damani et al. (2024) — and never includes information from the rollout itself. We write $\phi(x) \in \mathbb{R}^d$, $\|\phi(x)\|_2 \le 1$, for the feature map. The learner selects a level $k_t \in [K]$, the model generates one rollout under budget $b_{k_t}$, and the learner pays $c_{k_t}$ and observes the feedback described in §I.2.

**Quality, cost, utility.** Let $q(x, k) \in [0, 1]$ be the expected quality of a rollout for query $x$ under budget $b_k$ (for verifiable tasks, the probability that the extracted answer is correct), and let $Q_t(k) \in [0, 1]$ denote the realized quality at round $t$. Costs $c_k$ are *known, deterministic and strictly increasing*, $c_1 < \cdots < c_K$; we normalize $c_k = b_k / b_K \in (0, 1]$. For a cost-sensitivity parameter $\lambda \ge 0$ the utility of level $k$ is

$$u_\lambda(x, k) \;=\; q(x, k) \;-\; \lambda\, c_k, \tag{1}$$

and the per-context optimal level and its value are

$$k^\star_\lambda(x) \;\in\; \arg\max_{k \in [K]} u_\lambda(x, k), \qquad u^\star_\lambda(x) \;=\; \max_{k \in [K]} u_\lambda(x, k). \tag{2}$$

The scalarization (1) is the standard cost-aware reward used in LLM routing (Sridhar et al., 2026); $\lambda$ is set by the deployer and traces out an accuracy–cost frontier rather than being learned.

**Definition 1** (Contextual regret)**.** The regret of an algorithm $\mathcal{A}$ playing $k_1, \dots, k_T$ is

$$R_T(\mathcal{A}) \;=\; \sum_{t=1}^{T} \Big( u^\star_\lambda(x_t) - u_\lambda(x_t, k_t) \Big), \tag{3}$$

measured against the *per-context* optimum (2), not the best fixed level. The comparator is therefore the cheapest-adequate-budget oracle, which is the object our pilot measures.

*Remark* 2 (Token-indexed regret)*.* Because level $k$ consumes $b_k$ tokens, rounds are not the only meaningful clock. We write $\Lambda_T = \sum_{t \le T} b_{k_t}$ for tokens spent and, where relevant, report regret against $\Lambda$ as in the multi-fidelity bandit of Kandasamy et al. (2016). Round-indexed and token-indexed regret coincide up to the factor $b_K / b_1$, so all rates below are stated in rounds.

### I.2 Truncation feedback and its nesting

The mechanism we exploit is not our observation. Liao et al. (2025) truncate a chain of thought at multiple prefix depths, pay the thinking cost once, and harvest an answer at every depth — their cost model $B(n, m, H) = n[C_{\text{think}} + mHC_{\text{sol}}]$ carries no factor of $H$ on the thinking term. Qi et al. (2025) independently truncate one thinking process at several sampled budgets and extract a verifiable reward at each. Neither work selects a budget online or proves anything about regret. Our contribution starts one step later: we observe that this known mechanism is precisely a *one-sided, nested* observation structure, and we ask what an online allocator can prove given it.

Formally, let $\tau_j(\cdot)$ denote truncation of a token sequence to its first $b_j$ tokens followed by answer extraction. Playing level $k$ produces a rollout $\rho_t$ under budget $b_{k_t}$, and the learner forms

$$\widetilde{Y}_t(j) \;=\; \text{score}\big(\tau_j(\rho_t)\big), \qquad j \le k_t, \tag{4}$$

so that one paid rollout yields observations at every level $j \le k_t$ and none at $j > k_t$. We call $\widetilde{Y}_t(j)$ for $j < k_t$ the *truncated channel* and $Y_t(k_t) := Q_t(k_t)$, the score of the untruncated rollout, the *native channel*.

**Definition 3** (Nested feedback)**.** The induced feedback graph $G$ on $[K]$ has $k \to j$ for all $j \le k$: a complete lower-triangular (transitively closed) graph. Crucially, all $k$ observations arrive from *one* paid generation, whereas in the multi-fidelity bandit (Kandasamy et al., 2016) each fidelity is a separately paid pull. We refer to this as *nesting*, and it is the one structural property of compute ladders not present in the classical models.

*Remark* 4 (What nesting is worth, and what it is not worth)*.* $G$ is strongly observable with independence number $\alpha(G) = 1$ and domination number 1. By the trichotomy of Alon et al. (2015), the minimax round-indexed regret for such a graph is $\widetilde{\Theta}(\sqrt{\alpha T}) = \widetilde{\Theta}(\sqrt{T})$, with no dependence on $K$.

We state this plainly because it disciplines the rest of the paper: **a $\widetilde{O}(\sqrt{T})$ bound on the ladder is a corollary of a 2015 theorem and is not a contribution of this work.** Moreover Lykouris et al. (2019) show that graph-blind UCB and Thompson sampling already attain graph-structured rates, so "our algorithm consults the graph" is not by itself an advance. What the generic results do *not* cover is that the truncated channel is biased (§I.3), which is where our analysis lives.

### I.3 Assumptions

We state each assumption separately and record whether our pilot ($n = 65$ GSM8K/MATH-500 problems, five levels, independent rollouts at every level) supports it, contradicts it, or leaves it open. No result in §K uses Assumption 1; it is used only by an optional accelerator.

**Assumption 1** ($\varepsilon$-monotonicity in expectation)**.** There is $\varepsilon \geq 0$ such that for all $x$ and $j < k$, $q(x, j) \leq q(x, k) + \varepsilon$.

*Measurement status: supported only in the weak, in-expectation form; the per-instance form is contradicted.* Pooled forced-answer accuracy is increasing in the mean over most of the ladder ($0.529, 0.690, 0.683, 0.707, 0.740$ at $b = 128, \ldots, 2048$), so the population version holds with a small $\varepsilon$. But $7/40$ GSM8K and $3/35$ MATH-500 problems ($17.5\%$ and $8.6\%$) flip from correct to wrong as the budget grows, so the per-instance statement $Q_t(j) \leq Q_t(k)$ is false, consistent with the budget-dependent rank reversals reported by de Souza & Panisson (2026). We adopt the in-expectation form, and §K shows even this is dispensable: our main bound holds with $\varepsilon$ arbitrary, and Proposition 14 exhibits linear regret for an algorithm that treats Assumption 1 as a hard constraint.

**Assumption 2** (Truncation fidelity: per-level two-sided bounded bias)**.** For every level $j$ there is $\delta_j \geq 0$ such that for all $x$ and all $k > j$,

$$\left| \mathbb{E}\big[\widetilde{Y}_t(j) \,\big|\, x_t = x,\, k_t = k\big] - q(x, j) \right| \;\leq\; \delta_j, \tag{5}$$

and the native channel is unbiased, $\mathbb{E}[Y_t(k) \mid x_t = x] = q(x, k)$. Conditional on the past, the centered observations are $\frac{1}{2}$-sub-Gaussian. We write $\delta_{\max} = \max_j \delta_j$ and emphasize that (5) assumes *no sign* for the bias.

*Measurement status: supported, and quantified per level.* We measured $\delta_j$ directly as the excess pairwise disagreement of a truncated-versus-native comparison over a same-arm decoding-noise control (two native rollouts at the same budget), using an exact binomial McNemar test on paired per-problem agreement, $n \approx 80$ per level:

| $b_j$ | 128 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|
| noise-floor disagreement | 0.113 | 0.088 | 0.101 | 0.038 | 0.078 |
| truncated-vs-native disagreement | 0.225 | 0.100 | 0.139 | 0.050 | 0.091 |
| excess $= \widehat{\delta}_j$ | **0.113** | 0.013 | 0.038 | 0.013 | 0.013 |
| McNemar $p$ | **0.035** | 1.00 | 0.45 | 1.00 | 1.00 |

Two features of this table drive the rest of the paper. First, $\delta_j$ is *decreasing in* $j$: truncation bias is statistically indistinguishable from decoding noise at every budget except the shortest, where it is significant ($p = 0.035$). Second, at $b = 2048$ the "truncated" arm *is* the untruncated trace, so its $0.013$ excess isolates extractor stochasticity and calibrates the decomposition — the $0.013$–$0.038$ values above it are within that floor. We therefore model $\delta_j$ as per-level and decreasing rather than as a single global constant.

*The bias is not one-signed.* It is tempting to assume truncation is systematically pessimistic, since a trace cut mid-derivation has no answer to extract, and the budget-conditioning results of Li et al. (2025) make some distortion inevitable. A mechanism probe of the $41$ disagreeing pairs rejects this: $22$ had the truncated arm correct and the native arm wrong, $17$ the reverse, and only $2$ were cut mid-derivation with nothing extractable. Short native rollouts fail about as often as truncated prefixes, frequently by not committing to an answer within the cap. Disagreement is therefore closer to symmetric churn than to one-sided bias, which is why (5) is stated two-sided and why our confidence intervals inflate on both sides (§J). Directionality as a function of the gap $k - j$ is **unmeasured**; the $b = 128$ result is consistent with bias growing in $k - j$ but only one source budget was tested, so we flag this as suggestive, not established. All measurements are single-model.

**Assumption 3** (Realizability)**.** For each level $k$ there is $\theta_k^\star \in \mathbb{R}^d$ with $\|\theta_k^\star\|_2 \le S$ and $q(x,k) = \langle \phi(x), \theta_k^\star \rangle$ for all $x$.

*Pilot status: open — assumed, not tested.* With $n = 65$ problems we cannot fit or falsify a per-level linear quality model, and we do not claim to. 3 is the standard realizability condition of Chu et al. (2011) and Abbasi-Yadkori et al. (2011); misspecification of size $\eta$ adds the usual $O(\eta\sqrt{d}\,T)$ term, which is additive in $\delta_{\max}$'s place and does not interact with our analysis. Nothing in §K requires the $\theta_k^\star$ to be related across $k$.

**Assumption 4** (Context diversity)**.** $\Sigma = \mathbb{E}_{x\sim\mathcal{P}_x}[\phi(x)\phi(x)^\top]$ satisfies $\lambda_{\min}(\Sigma) \ge \sigma^2 > 0$.

*Pilot status: open; a property of the query stream, not of the model.* 4 is used *only* for the pooled branch of Theorem 8, which converts the free observations into a $K$-independent term; the exact-channel branch never needs it. Where it fails, Theorem 8 still holds through its other branch.

# J ALGORITHM

## J.1 WHY TWO CHANNELS

The naive reading of Remark 4 is that since $\alpha(G) = 1$, we should pool every observation — native and truncated alike — into one estimator per level and inherit a $\sqrt{T}$ bound. This is unsound. Under Assumption 2 the truncated channel for level $j$ estimates $q(x,j) - \delta_{j,k}(x)$, not $q(x,j)$, and pooling drives the estimator to a biased limit no matter how many samples arrive: the bias does not average out, because it is present in every truncated sample. Proposition 7 makes this precise, and our simulation confirms the mechanism: as the bias is scaled above the measured $\delta_k$, naive pooling's regret rises monotonically while TRUNCUCB's stays flat (Table 10).

TRUNCUCB therefore maintains *two* estimators per level and never mixes their samples:

- the **native (exact) channel** $\widehat{\theta}_k^{\mathrm{ex}}$, fitted only on rounds where $k$ was *played*, hence unbiased but fed at most one sample per round;
- the **pooled (truncated) channel** $\widehat{\theta}_k^{\mathrm{po}}$, fitted on every observation of level $k$ — native and truncated — hence fed $m_k = \sum_{i\ge k} n_i$ samples but biased by at most $\delta_k$ in magnitude, with *unknown sign*.

Since the bias has no known sign (our measurement finds it bidirectional), the pooled channel yields a valid interval only after *symmetric* inflation, by $\delta_k\sqrt{m_k}\|\phi_t\|_{(V_k^{\mathrm{po}})^{-1}}$ (Lemma 18); no free lower bound is available. Intersecting the two intervals then yields a confidence set never wider than either channel alone, which is the mechanism behind the adaptive minimum in Theorem 8. Because the inflation grows as $\sqrt{m_k}$ while the width shrinks as $m_k^{-1/2}$, the pooled interval is *not* monotonically improving in the amount of free data — the point made precise in Theorem 8(c).

## J.2 CONFIDENCE SETS

Let $n_k(t)$ be the number of rounds level $k$ was played and $m_k(t)$ the number of observations of level $k$ from either channel. With regularizer $\varrho > 0$ define the two Gram matrices and ridge solutions

$$V_k^{\mathrm{ex}}(t) = \varrho I + \sum_{s<t:\, k_s=k} \phi_s\phi_s^\top, \qquad \widehat{\theta}_k^{\mathrm{ex}}(t) = V_k^{\mathrm{ex}}(t)^{-1} \sum_{s<t:\, k_s=k} \phi_s\, Y_s(k), \tag{6}$$

$$V_k^{\mathrm{po}}(t) = \varrho I + \sum_{s<t:\, k_s\ge k} \phi_s\phi_s^\top, \qquad \widehat{\theta}_k^{\mathrm{po}}(t) = V_k^{\mathrm{po}}(t)^{-1} \sum_{s<t:\, k_s\ge k} \phi_s\, \widetilde{Y}_s(k), \tag{7}$$

where $\phi_s = \phi(x_s)$ and $\widetilde{Y}_s(k) = Y_s(k)$ when $k_s = k$. Write $\|\phi\|_{A^{-1}} = \sqrt{\phi^\top A^{-1}\phi}$ and let

$$\beta_n \;=\; \tfrac{1}{2}\sqrt{2\log(1/\delta) + d\log\bigl(1 + n/(d\varrho)\bigr)} \;+\; \sqrt{\varrho}\,S \tag{8}$$

**Algorithm 1** TRUNCUCB: contextual allocation over a nested compute ladder

**Require:** budgets $b_1 < \cdots < b_K$, costs $c_k = b_k/b_K$, weight $\lambda$, **per-level bias bounds** $\delta_1, \ldots, \delta_K$ (measured), confidence $\delta$, regularizer $\varrho$, monotone mode $\in \{\text{OFF}, \text{GUARDED}\}$, slack $\varepsilon$

$V_k^{\mathrm{ex}}, V_k^{\mathrm{po}} \leftarrow \varrho I$; $\widehat{\theta}_k^{\mathrm{ex}}, \widehat{\theta}_k^{\mathrm{po}} \leftarrow 0$; $\text{VIOL}_k \leftarrow \mathsf{false}$ for all $k \in [K]$
**for** $t = 1, \ldots, T$ **do**
  observe $x_t$; $\phi_t \leftarrow \phi(x_t)$
  $\tau_t \leftarrow \sigma\beta_t\sqrt{dKL/t}$; $\mathcal{T}_t \leftarrow \{k : \delta_k \le \tau_t\}$ ▷ trust set, Thm 8(b)
  compute $\mathrm{LCB}_k, \mathrm{UCB}_k$ by (8)–(11) for all $k$, using the pooled channel only for $k \in \mathcal{T}_t$
  **if** mode = GUARDED **then** ▷ optional; uses 1
    **for** $k = 2, \ldots, K$ **do**
      **if** $\neg\text{VIOL}_k$ **then** $\mathrm{LCB}_k \leftarrow \max\{\mathrm{LCB}_k,\ \mathrm{LCB}_{k-1} - \varepsilon\}$
      **end if**
    **end for**
  **end if**
  $\ell_t \leftarrow \max_{j\in[K]}\big(\mathrm{LCB}_j - \lambda c_j\big)$ ▷ best certified utility
  $\mathcal{W}_t \leftarrow \{k : 1 - \lambda c_k \ge \ell_t\} \cap \{k : \mathrm{UCB}_k - \lambda c_k \ge \ell_t\}$ ▷ cost domination (12)
  **if** $\mathcal{W}_t = \emptyset$ **then** $\mathcal{W}_t \leftarrow [K]$
  $k_t \leftarrow \arg\max_{k\in\mathcal{W}_t}\big(\mathrm{UCB}_k - \lambda c_k\big)$, ties broken toward *smaller* $b_k$ ▷ optimism, cheap-first
  generate one rollout $\rho_t$ under budget $b_{k_t}$; pay $c_{k_t}$
  $Y_t(k_t) \leftarrow \mathrm{score}(\rho_t)$ ▷ native: exact
  $\widetilde{Y}_t(j) \leftarrow \mathrm{score}(\tau_j(\rho_t))$ for $j < k_t$ ▷ truncated: free, biased
  **update exact:** $V_{k_t}^{\mathrm{ex}} \mathrel{+}= \phi_t\phi_t^\top$, $\widehat{\theta}_{k_t}^{\mathrm{ex}} \leftarrow$ ridge solution with $(\phi_t, Y_t(k_t))$
  **update pooled:** for $j \le k_t$ with $j \in \mathcal{T}_t$: $V_j^{\mathrm{po}} \mathrel{+}= \phi_t\phi_t^\top$, $\widehat{\theta}_j^{\mathrm{po}} \leftarrow$ ridge solution with $(\phi_t, \widetilde{Y}_t(j))$ ▷ untrusted levels discard the free sample
  **if** mode = GUARDED **then and** $\max_{j<k_t}\widetilde{Y}_t(j) > Y_t(k_t)$
    $\text{VIOL}_{k_t} \leftarrow \mathsf{true}$ ▷ observed prefix contradicts the monotone cone
  **end if**
**end for**

be the self-normalized radius of Abbasi-Yadkori et al. (2011). Define, for each level $k$,

$$\mathrm{L}_k^{\mathrm{ex}} = \langle\phi_t, \widehat{\theta}_k^{\mathrm{ex}}\rangle - \beta_{n_k}\|\phi_t\|_{(V_k^{\mathrm{ex}})^{-1}}, \qquad \mathrm{U}_k^{\mathrm{ex}} = \langle\phi_t, \widehat{\theta}_k^{\mathrm{ex}}\rangle + \beta_{n_k}\|\phi_t\|_{(V_k^{\mathrm{ex}})^{-1}}, \tag{9}$$

$$\mathrm{L}_k^{\mathrm{po}} = \langle\phi_t, \widehat{\theta}_k^{\mathrm{po}}\rangle - \big(\beta_{m_k} + \delta_k\sqrt{m_k}\big)\|\phi_t\|_{(V_k^{\mathrm{po}})^{-1}}, \quad \mathrm{U}_k^{\mathrm{po}} = \langle\phi_t, \widehat{\theta}_k^{\mathrm{po}}\rangle + \big(\beta_{m_k} + \delta_k\sqrt{m_k}\big)\|\phi_t\|_{(V_k^{\mathrm{po}})^{-1}}, \tag{10}$$

the inflation in (10) being symmetric because Assumption 2 bounds $|\cdot|$ without fixing a sign. The operative interval is the intersection

$$\mathrm{LCB}_k(t) = \max\{\mathrm{L}_k^{\mathrm{ex}}, \mathrm{L}_k^{\mathrm{po}}, 0\}, \qquad \mathrm{UCB}_k(t) = \min\{\mathrm{U}_k^{\mathrm{ex}}, \mathrm{U}_k^{\mathrm{po}}, 1\}. \tag{11}$$

### J.3 Cost domination: pruning that needs no assumption

Because $q \le 1$ and costs are known, level $k$ cannot beat a level already certified at utility $\mathrm{LCB}_j - \lambda c_j$ whenever $1 - \lambda c_k < \max_j(\mathrm{LCB}_j - \lambda c_j)$. This test uses *no* data about $k$ and *no* monotonicity: it follows from the known cost and the range of $q$. It defines the live set

$$\mathcal{W}_t \;=\; \Big\{k \in [K] \;:\; 1 - \lambda c_k \;\ge\; \max_j\big(\mathrm{LCB}_j(t) - \lambda c_j\big)\Big\}, \tag{12}$$

and Theorem 12 bounds regret by $|\mathcal{W}|$ rather than $K$. Expensive levels are eliminated permanently once any cheap level is certified good, which is what makes the algorithm cheap in tokens as well as in regret.

### J.4 The algorithm

### J.5 Reading the algorithm

**Exploration.** Optimism only, via (11); there is no forced exploration and no explicit exploration schedule. Levels are eliminated by (12) and by the intersection's upper bounds. The cheap-first tie-

break matters in practice: when two levels are statistically indistinguishable it spends fewer tokens, which is the behaviour the cheapest-correct oracle exhibits in our pilot.

**How the bias is handled.** The per-level bounds $\delta_k$ enter in exactly one place: the pooled interval (10), through the *symmetric* inflation $\delta_k\sqrt{m_k}\|\phi_t\|_{(V_k^{\mathrm{Po}})^{-1}}$ justified by Lemma 18. The inflation is symmetric because our measurement finds truncation disagreement bidirectional rather than systematically pessimistic, so no sign may be assumed; and it is not a bare $\delta_k$ because a systematic shift in the regression targets is amplified along poorly-conditioned directions of the design (Remark 19). Crucially the term does *not* vanish as $m_k$ grows: under Assumption 4 it equals the constant $\delta_k\sqrt{2}/\sigma$, so the free channel's precision saturates at $n_k^\star = 2\beta_T^2/\delta_k^2$ (Theorem 8(c)). Free observations past $n_k^\star$ are worthless for estimation, which is why they cannot substitute for native pulls.

**The trust set, and its empirically-checkable prescription.** Line 4 admits level $k$'s pooled channel only if $\delta_k \le \tau_t := \sigma\beta_t\sqrt{dKL/t}$, the threshold at which Theorem 8(b) says the pooled term becomes the smaller one. Because $\tau_t$ shrinks in $t$, levels are dropped from the trust set as the horizon lengthens, never added — the free channel is most useful early. Substituting our measured $\delta_k$ the rule trusts $b \ge 256$ and rejects $b = 128$, matching the only rung at which the bias is statistically significant. We report this as a prediction the measurement corroborates, while noting the converse honestly: at the measured $\delta_k$ the gated and ungated variants are statistically indistinguishable, so the rule is a safeguard for larger biases rather than a demonstrated win.

**What the free channel actually buys.** Three things, and we are explicit that statistical efficiency is not one of them at our measured biases (see Remark 11): (i) *token cost* — covering the ladder costs $b_K$ nested rather than $\sum_k b_k$ separately paid, a factor $\rho_K = 1.94$ here (Proposition 6); (ii) *cost-domination pruning* — the lower bounds $\mathrm{L}_k$ the free channel supplies are what make the test of Theorem 12 fire without spending a pull; (iii) *safety* — the intersection (11) always contains the native interval, so admitting the channel cannot invalidate the bound.

**What happens when monotonicity is violated.** In mode OFF — the default, and the mode Theorem 8 analyzes — nothing happens, because Assumption 1 is never used. Mode GUARDED is an accelerator: it propagates *lower* bounds upward, so a violation can only cost optimism (an underestimate of a good level), never soundness of the pruning rule, and the VIOL flag disables propagation for a level as soon as an observed prefix contradicts the cone. We deliberately do *not* offer the symmetric rule that propagates *upper* bounds downward: that rule prunes expensive levels on the strength of cheap ones, and Proposition 14 shows it incurs $\Omega(T)$ regret on a ladder that violates Assumption 1 by $q_1 - q_2 = 0.55$ — a violation our measured per-instance non-monotonicity rates (17.5% GSM8K, 8.6% MATH-500) show is not hypothetical. In simulation at $T = 6\times10^4$ the capped variant plays the wrong level on 65.1% of rounds with doubling ratio $R(T)/R(T/2) = 3.33$, indicating linear growth, against 1.03 for the default on the same instance.

**Complexity.** Each round costs $O(Kd^2)$ time with rank-one (Sherman–Morrison) updates and $O(Kd^2)$ memory, independent of $T$. The truncated scores require $K$ answer extractions per round but *no* additional generation, which is the entire economy of the method: the marginal inference cost of the free channel is the cost of re-running the verifier on prefixes.

## K What nested truncation feedback does and does not buy

We begin by conceding the result a reviewer will reach for first, because our contribution only makes sense once it is out of the way.

**Proposition 5** (The rate is not ours)**.** *The nested feedback graph $G$ of Definition 3 is strongly observable with $\alpha(G) = 1$. Consequently, in the adversarial round-indexed setting, the minimax regret over $G$ is $\widetilde{\Theta}\big(\sqrt{\alpha(G)\,T}\big) = \widetilde{\Theta}(\sqrt{T})$, independent of $K$, by the trichotomy of Alon et al. (2015, Thm. 1).*

Proposition 5 is a corollary of a 2015 theorem, obtained by computing one graph invariant. **We claim no credit for a $\widetilde{O}(\sqrt{T})$ rate on compute ladders, and we state explicitly that any such bound**

**— including the one our own algorithm satisfies — is subsumed.** Two further reductions deserve the same treatment. Since $G$ is transitively closed, the contextual analysis of Wen et al. (2024) — which gives lower bounds interpolating from the independence number to the MAS number and near-optimal algorithms for transitively closed graphs — applies to the ladder as well, and we do not claim novelty against it for the unbiased case. We also checked the closest monotone-feedback framework directly, since it is sometimes described in secondary sources as covering censored or truncated observations. Agarwal et al. (2023) obtain $\widetilde{O}(\sqrt{T})$ regret for monotone stochastic problems under semi-bandit feedback, but **their framework does not apply to a compute ladder: their monotonicity is monotonicity of the offline optimum under stochastic dominance of $n$ *independent* item distributions, whereas our $K$ levels are dependent views of a single latent trace, and their setting has no context.** (We verified by full-text search that the terms "censored", "truncated", and "contextual" do not occur in that paper.) We cite it as the nearest monotone-semi-bandit relative, not as a subsuming result. The same caution applies to the algorithmic side: since Lykouris et al. (2019) show that graph-blind UCB and Thompson sampling already attain graph-structured rates, an ablation of "graph-aware versus graph-blind" is not evidence of a contribution either. Our claims are the three that follow: nesting is worth a bounded constant and never a rate (§K.1); the generic machinery is *unsound* here because the free channel is biased, and the correct bound is an adaptive minimum (§K.2); and the resulting bias floor is information-theoretic rather than an artifact (§K.4).

### K.1 Nesting buys a bounded constant, not a rate

Nesting (Definition 3) is the one structural feature separating compute ladders from the multi-fidelity bandit of Kandasamy et al. (2016), where each fidelity is a separately paid pull. It is natural to hope this buys a rate improvement. It does not.

**Proposition 6** (Token value of nesting)**.** *To obtain one observation at every level $j \in [K]$, a learner with nested feedback pays $b_K$ tokens (one rollout at the top level), whereas a learner that must pay per level pays $\sum_{k=1}^{K} b_k$. The saving is exactly*

$$\rho_K \;=\; \frac{\sum_{k=1}^{K} b_k}{b_K}. \tag{13}$$

*For a geometric ladder $b_k = b_1 r^{k-1}$ with $r > 1$, $\rho_K = \frac{r^K - 1}{r^{K-1}(r-1)} < \frac{r}{r-1}$ uniformly in $K$; for the doubling ladder $r = 2$ this is $\rho_K < 2$. For a linear ladder $b_k = k b_1$, $\rho_K = (K+1)/2$.*

For our pilot ladder $\{128, \ldots, 2048\}$, $\rho_K = 1.94$: nesting saves slightly under a factor of two in tokens, and $\rho_K$ is bounded no matter how finely the ladder is subdivided. *This is a constant-factor statement, and we do not dress it as anything more.* Its practical force is that the constant is paid on *every* round, and it is bought at literally zero marginal generation cost — but it cannot change an exponent, and any claim that nesting improves the regret *rate* over MF-UCB would be false.

### K.2 The main bound: soundness under a biased free channel

The substantive point is that the generic machinery does not apply. Standard graph-feedback analyses assume side observations are unbiased draws of the neighbour's reward. Under Assumption 2 they are not, and the consequence is not a worse constant but a *qualitatively wrong* algorithm.

**Proposition 7** (Naive pooling is inconsistent)**.** *Consider any algorithm that pools truncated and native observations of a level into a single unweighted least-squares estimator, treating both as unbiased (the literal reading of the $\alpha(G) = 1$ reduction). There exist instances satisfying Assumptions 2–4 with bias exactly $\delta_{\max}$ on which its regret is $\Omega(\delta_{\max} T)$, i.e. linear.*

The reason is that the pooled estimator converges to $\theta_k^\star - \delta_{\max} \Sigma^{-1} \mathbb{E}[\phi]$-type limits, and a level whose true utility advantage is smaller than $\delta_{\max}$ is ranked incorrectly forever, no matter how much data arrives. Our simulation reproduces the mechanism: as the bias is scaled up from the measured $\delta_k$, the naive learner's regret and its empirical growth exponent both increase, while TruncUCB's stay flat (Table 10).

The correct guarantee therefore cannot be a single $\sqrt{T}$ expression: it must degrade to the bandit-feedback rate as $\delta_{\max}$ grows. This is what the two-channel construction of §J achieves.

**Theorem 8** (Adaptive two-channel regret with per-level bias)**.** *Let $L = \log\big(1 + T/(d\varrho)\big)$, let $\beta_T$ be as in* (8)*, and let $\mathcal{T} \subseteq [K]$ be any* trust set *of levels whose pooled channel the algorithm admits. Under Assumptions 2 and 3, and additionally 4 for the pooled terms,* TRUNCUCB *in mode* OFF *satisfies, with probability at least $1 - \delta$,*

$$R_T \;\le\; 2 \min_{\mathcal{T} \subseteq [K]} \Big\{ \underbrace{2\beta_T \sqrt{2dTL}\,\big|[K] \setminus \mathcal{T}\big|^{1/2}}_{\text{levels on the native channel}} \;+\; \sum_{k \in \mathcal{T}} \underbrace{\Big( \tfrac{2\beta_T}{\sqrt{K}} \sqrt{\tfrac{2T}{\sigma^2}} + \tfrac{\delta_k T \sqrt{2}}{K\sigma} \Big)}_{\text{level } k \text{ on the free channel}} \Big\}, \tag{14}$$

*and in particular $R_T \le 2\min\{B_{\mathrm{ex}}(T), B_{\mathrm{po}}(T)\}$ with $B_{\mathrm{ex}}(T) = 2\beta_T\sqrt{2dKTL}$ (take $\mathcal{T} = \emptyset$) and $B_{\mathrm{po}}(T) = 2\beta_T\sqrt{2T/\sigma^2} + \delta_{\max} T\sqrt{2}/\sigma$ (take $\mathcal{T} = [K]$). No monotonicity assumption is used. Consequently:*

*(a) $R_T = \widetilde{O}\big(\beta_T\sqrt{dKT}\big)$ always, matching LinUCB over $K$ arms — admitting the free channel never hurts, since the minimum includes $\mathcal{T} = \emptyset$;*

*(b) level $k$'s pooled term is the smaller one precisely when $\delta_k \lesssim \sigma\beta_T\sqrt{dKL/T}$, which is the **trust rule**: admit $k$ to $\mathcal{T}$ iff its measured $\delta_k$ clears that threshold;*

*(c) equivalently, level $k$'s pooled channel stops improving once it has*

$$n_k \;\gtrsim\; n_k^\star \;=\; 2\beta_T^2/\delta_k^2 \tag{15}$$

*observations: beyond $n_k^\star$ the interval is pinned at its bias floor $\delta_k\sqrt{2}/\sigma$ and additional free data cannot tighten it at all.*

### K.3 THE BOUND'S TRUST ORDERING REPRODUCES THE MEASURED BIAS ORDERING

Theorem 8(b) is a statement about which rungs of the ladder deserve to be believed, and it can be checked against a measurement that was made without reference to it. We state the comparison as a labelled result because it is the one place where our theory and our experiments corroborate each other, and we are explicit about the direction of inference.

**Proposition 9** (Trust ordering)**.** *Admit level $k$'s free channel iff $\delta_k \le \tau_T := \sigma\beta_T\sqrt{dKL/T}$ (Theorem 8(b)). Since $\tau_T$ is strictly decreasing in $T$, levels leave the trust set in* decreasing order of $\delta_k$*, and the first to leave is* $\arg\max_k \delta_k$*. At the measured $\widehat{\delta} = (0.1125,\, 0.0125,\, 0.0380,\, 0.0125,\, 0.0130)$ with $d = 6$, $K = 5$, $\sigma^2 = 0.098$, $\varrho = 1$, $\delta_{\mathrm{conf}} = 0.05$, the order of departure is $b = 128$ (at $T \approx 5.2 \times 10^4$), then $b = 512$ (at $T \approx 7 \times 10^5$), then the remaining rungs.*

*Proof.* $\tau_T \propto \beta_T\sqrt{L/T}$ with $\beta_T, L$ growing only logarithmically, so $\tau_T \to 0$ monotonically for $T$ past the burn-in; the admission test $\delta_k \le \tau_T$ therefore fails first for the largest $\delta_k$. The two numerical thresholds are the roots of $\tau_T = \delta_k$ for $\delta_k \in \{0.1125, 0.0380\}$, computed in `verify_theory.py`. □

**What is and is not predicted.** The *ordering* is a derivation: it follows from the bound, uses no experimental input beyond the values $\widehat{\delta_k}$, and identifies $b = 128$ as the rung to distrust first. The measurement, performed independently, finds $b = 128$ to be the *only* rung whose truncation bias is statistically distinguishable from decoding noise ($p = 0.035$; all others $p \ge 0.45$). Theory and experiment therefore agree on which rung is untrustworthy, by two unrelated routes.

We are careful not to overclaim. The *absolute* threshold depends on free constants we fixed rather than derived — $\varrho = 1$, the confidence level, and the $O(1)$ factors absorbed into $\beta_T$ — so at our verification horizon $T = 4 \times 10^4$ we have $\tau_T = 0.125 > 0.1125$ and the rule still admits every level; $b = 128$ is excluded only for $T \gtrsim 5.2 \times 10^4$. The prediction we claim is the ordering and the identity of the first rung to be dropped, not a horizon-free cutoff, and the agreement is a consistency check rather than a fit: no constant was tuned to make $b = 128$ come out first.

*Remark* 10 (Saturation). Claim (c) is the sharpest statement in this paper and the one most specific to truncation feedback. The pooled channel's stochastic width shrinks as $n_k^{-1/2}$ while its bias contribution is *constant* in $n_k$: by Lemma 18 the bias displacement is $\delta_k\sqrt{n_k}\|\phi\|_{V^{-1}}$, and under 4 $\|\phi\|_{V^{-1}} \asymp \sqrt{2/(\sigma^2 n_k)}$, so the $\sqrt{n_k}$ cancels exactly. Free observations therefore buy precision only

up to $n_k^\star$ and nothing after. Substituting the measured $\delta_k$ into (15) at $T = 4 \times 10^4$, $d = 6$, $K = 5$ (so $\beta_T = 4.90$) and the measured $\sigma^2 = 0.098$ gives

| $b_k$ | 128 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|
| $\widehat{\delta_k}$ | 0.1125 | 0.0125 | 0.0380 | 0.0125 | 0.0130 |
| $n_k^\star$ | $3.8 \times 10^3$ | $3.1 \times 10^5$ | $3.3 \times 10^4$ | $3.1 \times 10^5$ | $2.8 \times 10^5$ |
| bias floor $\delta_k\sqrt{2}/\sigma$ | 0.508 | 0.056 | 0.171 | 0.056 | 0.059 |

— a spread of nearly two orders of magnitude between the shortest rung ($n^\star = 3.8e{+}03$ at $b = 128$) and the rest of the ladder ($n^\star \geq 3.3e + 04$, and $\approx 3 \times 10^5$ at $b \in \{256, 1024, 2048\}$). The consequences for which rungs to believe are worked through in §K.3.

The $K$-freeness in (b) comes from the nesting, through a telescoping argument that is the technical heart of the proof: writing $m_j = \sum_{i \geq j} n_i$ for the number of observations truncation delivers to level $j$,

$$\sum_{j=1}^{K} \frac{n_j}{\sqrt{m_j}} \;=\; \sum_{j=1}^{K} \frac{m_j - m_{j+1}}{\sqrt{m_j}} \;\leq\; \int_0^T \frac{du}{\sqrt{u}} \;=\; 2\sqrt{T}, \tag{16}$$

independent of $K$. The lower-triangular structure is exactly what makes $m_j$ decreasing in $j$ and hence makes (16) available.

*Remark* 11 (Honest comparison). Theorem 8 is **not** tighter in rate than Proposition 5: at $\delta_{\max} = 0$ both give $\widetilde{O}(\sqrt{T})$, and the $\alpha$-based result achieves that in the harder adversarial model. Our bound's value is different in kind. (i) It is *applicable*: the $\alpha$-based bounds assume unbiased side observations and, by Proposition 7, the algorithm they license is inconsistent here, so invoking $\alpha = 1$ on a compute ladder is unsound rather than merely loose. (ii) It is *finite-horizon informative*: it identifies the regime $T \lesssim \sigma^2 \beta_T^2 dKL/\delta_{\max}^2$ where the free channel pays and the regime beyond it where it must be discounted — a distinction an asymptotic-in-$T$ rate cannot express. (iii) It degrades *gracefully*, recovering the bandit-feedback guarantee instead of failing, which is what makes the method deployable at an empirically measured $\delta_{\max}$ rather than an assumed $\delta_{\max} = 0$. We regard this as the honest position, and we do not claim a rate improvement.

Two further concessions belong here rather than in a rebuttal. First, at the *measured* $\delta_k$ the free channel does not reduce regret in our simulations: on 24 paired instances TRUNCUCB is slightly *worse* than using the native channel alone ($+18.4 \pm 6.3$ regret, $p = 0.0033$), because the bias inflation (18) costs more than the extra observations buy once $n_k$ exceeds $n_k^\star$. The free channel's value in this regime is therefore *token cost* (Proposition 6) and the Theorem 12 pruning, not statistical efficiency. Second, since $\alpha(G) = 1$ and Lykouris et al. (2019) show graph-blind UCB already attains graph-structured rates, we do not present "uses the feedback graph" as an advantage and our ablations should not be read that way.

**Theorem 12** (Cost-domination confinement)**.** *Let $\mathcal{W}(x) = \{k : 1 - \lambda c_k \geq u_\lambda^\star(x)\}$ be the set of levels not cost-dominated at $x$, and $W = \sup_x |\mathcal{W}(x)|$. Then* TRUNCUCB *plays outside $\mathcal{W}(x_t)$ at most $O\big(\beta_T^2 d \log T/\Delta_{\min}^2\big)$ times in total, and $K$ may be replaced by $W$ in $B_{\mathrm{ex}}(T)$ of* (14)*, where $\Delta_{\min}$ is the smallest positive utility gap. The test requires no data about the pruned levels and no monotonicity assumption.*

Theorem 12 is the practical counterpart of Proposition 6: as $\lambda$ grows, expensive levels are excluded on arithmetic alone and both regret and token spend fall. Our simulation confirms the mechanism, with mean tokens per query decreasing monotonically in $\lambda$ (Table 10).

### K.4 TIGHTNESS, AND WHAT BREAKS THE ASSUMPTIONS

The $\delta_{\max} T$ term in (14) is not an artifact of our analysis. It reflects a genuine identifiability failure.

**Proposition 13** (The bias floor is information-theoretic)**.** *Fix $v \in (0, 1)$ and $\delta_j \in (0, 1 - v)$. Consider two instances that agree on all costs and contexts, both admissible under Assumption 2 with the same bound $\delta_j$:*

$$\mathcal{I}_A : \; q_A(x, j) = v, \;\; \mathit{bias}_A = 0; \qquad \mathcal{I}_B : \; q_B(x, j) = v + \delta_j, \;\; \mathit{bias}_B = -\delta_j.$$

*Both induce* identical *laws for every truncated observation (total variation distance* $0$*), yet their optimal levels differ and any algorithm that uses only truncated observations of a level suffers per-round regret at least* $\delta_j/2$ *on one of them. Hence no algorithm can achieve a bound of the form* $o(\delta_j T)$ *using the free channel alone, and native pulls are necessary to break the tie.*

Proposition 13 explains why TRUNCUCB keeps a separate exact channel rather than de-biasing: the bias is not identifiable from truncated data, so it can only be bounded, never estimated away. This also distinguishes our setting from Kandasamy et al. (2016), whose fidelity gaps $\zeta^{(m)}$ are *known* and whose lowest fidelity is still an unbiased draw of its own distribution, and from Sridhar et al. (2026), whose surrogate bias $b_i$ is a constant per arm and hence estimable by de-biasing. Our $\delta_j$ bounds a bias of unknown sign that our measurement finds to be bidirectional, so there is no constant offset to remove.

**Proposition 14** (Monotonicity, taken as a hard constraint, is catastrophic)**.** *Consider* $K = 3$ *with* $q = (0.90, 0.35, 0.60)$*, costs* $c = (0.90, 0.95, 1.00)$ *and* $\lambda = 0.1$*, so that utilities are* $u = (0.81, 0.255, 0.50)$ *and* $k^\star = 1$*: quality is badly non-monotone in budget, violating Assumption 1 by* $q_1 - q_2 = 0.55$*. An algorithm that enforces* $\varepsilon$*-monotonicity by capping lower levels' upper confidence bounds via higher ones,* $\mathrm{U}_j \leftarrow \min\{\mathrm{U}_j, \mathrm{U}_{j+1} + \varepsilon\}$ *with* $\varepsilon$ *set from a measured violation rate, incurs* $\Omega(T)$ *regret, because the cap slanders the optimal level permanently.* TRUNCUCB *in mode* OFF *or* GUARDED *is unaffected.*

In simulation at $T = 6 \times 10^4$ the capped variant holds an *empty* confidence interval at the optimal level ($\mathrm{U}_1 = 0.578 < \mathrm{L}_1 = 0.784$), plays the suboptimal level 3 on $65.1\%$ of rounds, and incurs $22.4\times$ the regret of the monotonicity-free default ($12115$ against $540$), with doubling ratio $R(T)/R(T/2) = 3.33$ against the default's $1.03$. Since we measure per-instance violation rates of $17.5\%$ on GSM8K and $8.6\%$ on MATH-500, and de Souza & Panisson (2026) report budget-dependent rank reversals, an algorithm whose correctness rests on monotonicity is not deployable. This is why no monotonicity assumption appears in Theorem 8.

*Remark* 15 (Lower bounds and the remaining gap). Three lower bounds bracket our result. (i) $\Omega(\sqrt{T})$ holds already for $K = 2$ with full information, so the $\sqrt{T}$ in (14) is unimprovable in $T$. (ii) $\Omega(\delta_j T)$ for truncation-only algorithms, by Proposition 13. (iii) $\Omega(\sqrt{dT})$ for linear contextual bandits (Chu et al., 2011) applies to the native channel. What we do *not* have is a lower bound establishing that the *minimum* in (14) is the optimal interpolation between the two channels, nor one showing the $\sqrt{K}$ in $B_{\mathrm{ex}}$ is necessary given nesting; an algorithm that adaptively estimates $\delta_{j,k}(x)$ from the sparse native pulls might interpolate better. We state this as an open problem rather than asserting optimality.

## L PROOFS

Throughout, $\phi_t = \phi(x_t)$ with $\|\phi_t\|_2 \le 1$, $\varrho > 0$ is the ridge regularizer, $n_k(t)$ and $m_k(t)$ are as in §J, and $\beta_n$ is defined in (8). All quality values lie in $[0, 1]$, so rewards are $\frac{1}{2}$-sub-Gaussian after centering. We abbreviate $\|\phi\|_{A^{-1}} = \sqrt{\phi^\top A^{-1}\phi}$. We use the following two standard facts.

**Lemma 16** (Self-normalized concentration; Abbasi-Yadkori et al., 2011, Thm. 2)**.** *Let* $(\phi_s, \eta_s)_{s\ge1}$ *be adapted with* $\eta_s$ *conditionally* $R$*-sub-Gaussian and zero-mean,* $V_n = \varrho I + \sum_{s\le n} \phi_s \phi_s^\top$*, and* $\widehat{\theta}_n = V_n^{-1} \sum_{s\le n} \phi_s(\langle \phi_s, \theta^\star\rangle + \eta_s)$*. Then with probability at least* $1-\delta$*, simultaneously for all* $n \ge 0$*,* $\|\widehat{\theta}_n - \theta^\star\|_{V_n} \le R\sqrt{2\log(1/\delta) + d\log(1 + n/(d\varrho))} + \sqrt{\varrho}S = \beta_n$.

**Lemma 17** (Elliptical potential; Abbasi-Yadkori et al., 2011, Lem. 11)**.** *For any sequence* $\phi_1, \dots, \phi_n$ *with* $\|\phi_s\|_2 \le 1$ *and* $V_s = \varrho I + \sum_{r<s} \phi_r \phi_r^\top$*,* $\sum_{s=1}^n \|\phi_s\|^2_{V_s^{-1}} \le 2d\log\big(1 + n/(d\varrho)\big)$.

The next lemma is the one piece of machinery specific to our setting: it quantifies how a one-sided bounded bias in the regression targets displaces a ridge estimate. It licenses the inflated pooled upper bound (10), and it shows the inflation must scale as $\delta_{\max}\sqrt{m}\,\|\phi\|_{V^{-1}}$ rather than as a bare $\delta_{\max}$.

**Lemma 18** (Bias displacement of a ridge estimate)**.** *Let* $V = \varrho I + \sum_{s=1}^m \phi_s \phi_s^\top$ *with* $\|\phi_s\|_2 \le 1$*, and let*

$$\widehat{\theta} = V^{-1} \sum_{s=1}^m \phi_s\big(\langle \phi_s, \theta^\star\rangle - \delta_s + \eta_s\big), \qquad \widetilde{\theta} = V^{-1} \sum_{s=1}^m \phi_s\big(\langle \phi_s, \theta^\star\rangle + \eta_s\big),$$

Table 10: Numerical verification of every theoretical claim. All quantities are computed by `verify_theory.py` using the *theoretical* confidence radius (no tuned exploration constant), with $\sigma^2$ measured rather than assumed, on instances calibrated to the measured ladder accuracies, per-level $\delta_k$, and 13.3% non-monotonicity. Rows 6 and 9 are negative results we report rather than omit: at the measured biases the free channel does not improve statistical efficiency, and the trust rule is not empirically separable, so the free channel's value lies in token cost (Proposition 6) and cost-domination pruning (Theorem 12) instead. Every value comes from a single run of the rebuilt verification script (see the Reproducibility Statement).

| Check | Measured | Theory predicts |
|---|---|---|
| Empirical regret $\leq$ proven bound (all $T$) | 621 vs bound 53,145 | holds |
| Growth: doubling ratio $R(T)/R(T/2)$ | 1.34 | $\sqrt{T} \Rightarrow 1.41$ |
| Nesting saving $\rho_K$ (pilot ladder) | 1.94 | $< r/(r-1) = 2$ |
| Saturation $n_k^\star$: $b$=128 / $b$=256 | $3.8\times10^3$ / $3.1\times10^5$ | 2 orders apart |
| Identifiability: joint TV distance | 0 exactly | $= 0$ |
| Free channel improves regret at measured $\delta_k$? | $-18.4 \pm 6.3$, $p = 0.0033$ | **no** |
| Naive pooling as bias scales 0→8× | regret 554→704, slope 0.45→0.53 | degrades |
| TRUNCUCB as bias scales 0→8× | regret 597→601 | stable |
| Trust rule separable at measured $\delta_k$? | indistinguishable | **no** |
| First rung to leave the trust set | $b$=128 ($T \approx 5.2\times10^4$) | largest $\delta_k$ first |
| Hard $\varepsilon$-monotone capping | 22.4× regret, ratio 3.33 vs 1.03 | linear (fails) |
| Cost domination: tokens/query vs $\lambda$ | 1132→304 | falls monotonically |

*where* $0 \leq \delta_s \leq \delta_{\max}$*, so that* $\widehat{\theta} = \widetilde{\theta} - V^{-1}\sum_s \phi_s\delta_s$*. Then for every* $\phi$ *with* $\|\phi\|_2 \leq 1$*,*

$$\left|\langle \phi, \widetilde{\theta} - \widehat{\theta}\rangle\right| \;\leq\; \delta_{\max}\sqrt{m}\,\|\phi\|_{V^{-1}}. \tag{17}$$

*Proof.* Put $b = \sum_{s=1}^m \phi_s\delta_s$, so $\langle \phi, \widetilde{\theta} - \widehat{\theta}\rangle = \phi^\top V^{-1} b$. Since $V \succ 0$, the bilinear form $\langle u, v\rangle_{V^{-1}} = u^\top V^{-1} v$ is an inner product, and Cauchy–Schwarz gives $|\phi^\top V^{-1} b| \leq \|\phi\|_{V^{-1}}\|b\|_{V^{-1}}$.

To bound $\|b\|_{V^{-1}}$, let $\Phi \in \mathbb{R}^{m\times d}$ have rows $\phi_s^\top$ and let $\delta = (\delta_1, \ldots, \delta_m)^\top$, so $b = \Phi^\top\delta$ and $\|b\|_{V^{-1}}^2 = \delta^\top\Phi V^{-1}\Phi^\top\delta$. Let $\Phi = U\Sigma W^\top$ be a singular value decomposition with singular values $\sigma_i \geq 0$. Then $\Phi^\top\Phi = W\Sigma^2 W^\top$ and $V = \varrho I + W\Sigma^2 W^\top$, so

$$\Phi V^{-1}\Phi^\top \;=\; U\Sigma W^\top\big(W(\varrho I + \Sigma^2)W^\top\big)^{-1} W\Sigma U^\top \;=\; U\,\Sigma(\varrho I + \Sigma^2)^{-1}\Sigma\,U^\top,$$

a symmetric matrix whose eigenvalues are $\sigma_i^2/(\varrho + \sigma_i^2) \leq 1$. Hence $\|b\|_{V^{-1}}^2 \leq \|\delta\|_2^2 \leq m\delta_{\max}^2$, and $\|b\|_{V^{-1}} \leq \delta_{\max}\sqrt{m}$, which with Cauchy–Schwarz gives (17). □

*Remark* 19 (Why the inflation is not a bare $\delta_{\max}$). It is tempting to argue that since each target is displaced by at most $\delta_{\max}$, the estimate is displaced by at most $\delta_{\max}$. That is false in general: ridge regression can amplify a systematic target shift along poorly-conditioned directions, and (17) is the correct $\|\cdot\|_{V^{-1}}$-weighted statement. Under Assumption 4, however, $\lambda_{\min}(V) \geq \sigma^2 m/2$ for $m$ past the burn-in, so $\delta_{\max}\sqrt{m}\,\|\phi\|_{V^{-1}} \leq \delta_{\max}\sqrt{m}\cdot\sqrt{2/(\sigma^2 m)} = \delta_{\max}\sqrt{2}/\sigma$, a constant per round — which is what produces the $\delta_{\max}T\sqrt{2}/\sigma$ term in (14). Our implementation uses the computable form $(\beta_m + \delta_{\max}\sqrt{m})\|\phi\|_{V^{-1}}$ directly and therefore never needs to know $\sigma$.

### L.1 Proof of Proposition 5

*Proof.* Recall $G$ has vertex set $[K]$ and arc $k \to j$ iff $j \leq k$; in particular every vertex has a self-loop, so every vertex is strongly observable and $G$ is a strongly observable graph in the sense of Alon et al. (2015).

*Independence number.* Let $\{j, k\}$ with $j < k$ be any two distinct vertices. Since $j \leq k$, the arc $k \to j$ is present, so $\{j, k\}$ is not an independent set. Hence no independent set has size 2, and as singletons are independent, $\alpha(G) = 1$.

*Domination number.* The single vertex $K$ has out-neighbourhood $\{j : j \leq K\} = [K]$, so $\{K\}$ dominates $G$ and the domination number is 1.

*Transitivity.* If $k \to j$ and $j \to i$ then $i \le j \le k$, so $k \to i$; the arc relation is transitive, i.e. $G$ is transitively closed.

Applying Alon et al. (2015, Thm. 1) to the strongly observable graph $G$ gives minimax regret $\widetilde{\Theta}(\alpha^{1/2}T^{1/2}) = \widetilde{\Theta}(T^{1/2})$, with no $K$ dependence. □

### L.2 Proof of Proposition 6

*Proof.* Under nesting, one rollout at level $K$ costs $b_K$ tokens and, by (4), yields an observation at every $j \le K$, i.e. at all $K$ levels; no cheaper single action observes level $K$, since observing level $j$ requires generating at least $b_j$ tokens and $b_K$ is the largest. Without nesting each level must be generated separately, costing $\sum_{k=1}^{K} b_k$. The ratio is $\rho_K = \sum_k b_k / b_K$ by definition.

For $b_k = b_1 r^{k-1}$ with $r > 1$, $\sum_{k=1}^{K} b_k = b_1 \frac{r^K - 1}{r-1}$ and $b_K = b_1 r^{K-1}$, so

$$\rho_K = \frac{r^K - 1}{r^{K-1}(r-1)} = \frac{r}{r-1}\left(1 - r^{-K}\right) < \frac{r}{r-1},$$

uniformly in $K$, and $\rho_K \uparrow r/(r-1)$. For $r = 2$ the bound is 2. For $b_k = k b_1$, $\sum_k b_k = b_1 K(K+1)/2$ and $b_K = b_1 K$, so $\rho_K = (K+1)/2$. For the pilot ladder, $\rho_5 = (128 + 256 + 512 + 1024 + 2048)/2048 = 3968/2048 = 1.9375$. □

### L.3 Proof of Proposition 7

*Proof.* We exhibit an instance. Take $d = 1$ with $\phi(x) \equiv 1$ (so Assumptions 3 and 4 hold with $S = 1$, $\sigma^2 = 1$), $K = 2$, $\lambda = 0$, and quality values $q(1) = v + \frac{3}{4}\delta_{\max}$, $q(2) = v$ for some $v \in (0, 1 - \delta_{\max})$. Thus level 1 is optimal with utility gap $\Delta = \frac{3}{4}\delta_{\max} > 0$. Let the truncation bias be exactly $\delta_{\max}$: whenever level 2 is played, the truncated observation of level 1 has mean $q(1) - \delta_{\max} = v - \frac{1}{4}\delta_{\max}$.

Consider any algorithm that maintains, for each level $k$, a single unweighted least-squares estimator over all observations of $k$ (native and truncated alike), and selects greedily/optimistically with confidence widths that shrink to 0 as the count grows — this is the literal $\alpha(G) = 1$ reduction, which treats every in-neighbour observation as an unbiased sample.

Suppose for contradiction the algorithm plays level 2 only $o(T)$ times. Then it plays level 1 for $T - o(T)$ rounds, and since playing 1 reveals only level 1, level 2's estimator receives $o(T)$ samples; its confidence width does not vanish, so an optimistic rule keeps selecting 2 — contradiction unless 2 is played $\Omega(T)$ times. Hence level 2 is played $\Omega(T)$ times; write $n_2 = \Omega(T)$.

Each of those $n_2$ rounds contributes one truncated sample of level 1 with mean $v - \frac{1}{4}\delta_{\max}$, while native samples of level 1 have mean $v + \frac{3}{4}\delta_{\max}$. If level 1 is played $n_1$ times, the pooled estimator for level 1 concentrates (as $n_1 + n_2 \to \infty$) on

$$\widehat{q}(1) \;\longrightarrow\; \frac{n_1\left(v + \frac{3}{4}\delta_{\max}\right) + n_2\left(v - \frac{1}{4}\delta_{\max}\right)}{n_1 + n_2} \;=\; v + \delta_{\max}\,\frac{\frac{3}{4}n_1 - \frac{1}{4}n_2}{n_1 + n_2}.$$

Whenever $n_2 > 3n_1$, this limit is $< v = q(2)$, while level 2's estimator concentrates on $q(2) = v$ (level 2 is only ever observed natively, since no level above it is played). So the algorithm's ranking of the two levels is inverted relative to the truth, and it prefers level 2.

Preferring level 2 increases $n_2$ relative to $n_1$, which by the display above only strengthens the inversion: the set $\{n_2 > 3n_1\}$ is absorbing under the algorithm's own selection rule once entered, because every additional play of 2 increases $n_2$ and leaves $n_1$ fixed. As confidence widths shrink to 0, the algorithm therefore plays level 2 on all but $o(T)$ rounds, each incurring instantaneous regret $\Delta = \frac{3}{4}\delta_{\max}$. Hence $R_T \ge \frac{3}{4}\delta_{\max}\,(T - o(T)) = \Omega(\delta_{\max} T)$.

The construction extends to $K > 2$ and to non-degenerate contexts by embedding this pair as levels $\{1, 2\}$ and making all other levels cost-dominated (so Theorem 12 removes them), and by taking $\phi$ with $\lambda_{\min}(\Sigma) = \sigma^2 > 0$, which changes only constants. □

### L.4 Proof of Theorem 8

*Proof.* Condition on the event $\mathcal{E}$ that the conclusion of Lemma 16 holds simultaneously for all $2K$ estimators, each invoked at confidence $\delta/(2K)$; a union bound gives $\Pr[\mathcal{E}] \geq 1-\delta$. Absorbing the $\log(2K/\delta)$ into $\beta$ (which (8) does up to the stated constant) we write $\beta_n$ throughout. All statements below hold on $\mathcal{E}$.

*Step 1: validity of the intersected interval.* For the native channel, the noise $Y_s(k) - q(x_s,k)$ is zero-mean and $\frac{1}{2}$-sub-Gaussian, so Lemma 16 gives $|\langle \phi_t, \widehat{\theta}^{\mathrm{ex}}_k - \theta^\star_k\rangle| \leq \beta_{n_k}\|\phi_t\|_{(V^{\mathrm{ex}}_k)^{-1}}$ by Cauchy–Schwarz in the $V^{\mathrm{ex}}_k$ norm; hence $\mathrm{L}^{\mathrm{ex}}_k \leq q(x_t,k) \leq \mathrm{U}^{\mathrm{ex}}_k$.

For the pooled channel, decompose each sample as $\widetilde{Y}_s(k) = \langle \phi_s, \theta^\star_k\rangle - \delta_{k,k_s}(x_s) + \eta_s$ with $\eta_s$ zero-mean $\frac{1}{2}$-sub-Gaussian and $0 \leq \delta_{k,k_s}(x_s) \leq \delta_{\max}$ by Assumption 2 (with $\delta_{k,k} \equiv 0$ for native samples). By linearity of the ridge solution in its targets, $\widehat{\theta}^{\mathrm{po}}_k = \widetilde{\theta}_k - \xi_k$ where $\widetilde{\theta}_k$ solves the same problem with the *unbiased* targets $\langle \phi_s, \theta^\star_k\rangle + \eta_s$ and $\xi_k = (V^{\mathrm{po}}_k)^{-1}\sum_s \phi_s\, \delta_{k,k_s}(x_s)$.

Lemma 16 applies to $\widetilde{\theta}_k$ and gives $|\langle \phi_t, \widetilde{\theta}_k - \theta^\star_k\rangle| \leq \beta_{m_k}\|\phi_t\|_{(V^{\mathrm{po}}_k)^{-1}}$, while Lemma 18 applied with $m = m_k$ and bias bound $\delta_k$ gives $|\langle \phi_t, \xi_k\rangle| \leq \delta_k\sqrt{m_k}\,\|\phi_t\|_{(V^{\mathrm{po}}_k)^{-1}}$. Combining the two by the triangle inequality,

$$\left|\langle \phi_t, \widehat{\theta}^{\mathrm{po}}_k\rangle - q(x_t,k)\right| \;\leq\; \left(\beta_{m_k} + \delta_k\sqrt{m_k}\right)\|\phi_t\|_{(V^{\mathrm{po}}_k)^{-1}} \;=:\; \Gamma_k(t), \tag{18}$$

so with $\mathrm{L}^{\mathrm{po}}_k = \langle \phi_t, \widehat{\theta}^{\mathrm{po}}_k\rangle - \Gamma_k(t)$ and $\mathrm{U}^{\mathrm{po}}_k = \langle \phi_t, \widehat{\theta}^{\mathrm{po}}_k\rangle + \Gamma_k(t)$ we have $\mathrm{L}^{\mathrm{po}}_k \leq q(x_t,k) \leq \mathrm{U}^{\mathrm{po}}_k$.

Both bounds carry the bias inflation, symmetrically. We note explicitly that *no* one-sided sharpening is available: Lemma 18 bounds only the *magnitude* $|\langle \phi_t, \xi_k\rangle|$, and the sign of $\langle \phi_t, \xi_k\rangle = \sum_s (\phi_t^\top (V^{\mathrm{po}}_k)^{-1}\phi_s)\,\delta_s$ is not determined even when every $\delta_s \geq 0$, because the weights $\phi_t^\top (V^{\mathrm{po}}_k)^{-1}\phi_s$ can be negative. Independently of that algebraic point, our measurements reject a one-signed model empirically (bidirectional disagreement, 22 versus 17), so Assumption 2 is stated two-sided and the symmetric inflation (18) is the correct construction. Consequently $\mathrm{LCB}_k(t) \leq q(x_t,k) \leq \mathrm{UCB}_k(t)$ for every $k$ and $t$: the intersected interval (11) is valid, and by construction it is contained in each channel's interval.

*Step 2: optimism and the per-round regret decomposition.* By Step 1, $\mathrm{UCB}_k(t) - \lambda c_k \geq u_\lambda(x_t,k)$ for every $k$; in particular for $k = k^\star_\lambda(x_t)$. Also $\mathrm{LCB}_j - \lambda c_j \leq u_\lambda(x_t,j) \leq u^\star_\lambda(x_t) \leq 1 - \lambda c_{k^\star}$, so $k^\star_\lambda(x_t) \in \mathcal{W}_t$: the optimal level is never pruned (this proves the soundness half of Theorem 12). Since $k_t$ maximizes $\mathrm{UCB}_k - \lambda c_k$ over $\mathcal{W}_t$,

$$u^\star_\lambda(x_t) \;\leq\; \mathrm{UCB}_{k^\star}(t) - \lambda c_{k^\star} \;\leq\; \mathrm{UCB}_{k_t}(t) - \lambda c_{k_t},$$

whence the instantaneous regret obeys

$$u^\star_\lambda(x_t) - u_\lambda(x_t,k_t) \;\leq\; \mathrm{UCB}_{k_t}(t) - q(x_t,k_t) \;\leq\; \mathrm{UCB}_{k_t}(t) - \mathrm{LCB}_{k_t}(t) \;=\; w_t, \tag{19}$$

the width of the intersected interval at the played level. Because the intersection is contained in both channels' intervals, $w_t \leq \min\{w^{\mathrm{ex}}_t, w^{\mathrm{po}}_t\}$ where $w^{\mathrm{ex}}_t = 2\beta_{n_{k_t}}\|\phi_t\|_{(V^{\mathrm{ex}}_{k_t})^{-1}}$ and $w^{\mathrm{po}}_t = 2\beta_{m_{k_t}}\|\phi_t\|_{(V^{\mathrm{po}}_{k_t})^{-1}} + \delta_{\max}$. Summing (19) and bounding the sum by each branch separately gives $R_T \leq \min\{\sum_t w^{\mathrm{ex}}_t,\ \sum_t w^{\mathrm{po}}_t\}$, so it suffices to bound each sum.

*Step 3a: the native branch.* Group rounds by the played level and apply Cauchy–Schwarz then Lemma 17 within each group:

$$\sum_{t=1}^{T} w^{\mathrm{ex}}_t \;=\; 2\sum_{k=1}^{K}\sum_{t:k_t=k} \beta_{n_k(t)}\|\phi_t\|_{(V^{\mathrm{ex}}_k)^{-1}} \;\leq\; 2\beta_T \sum_{k=1}^{K}\sqrt{n_k(T)}\sqrt{\sum_{t:k_t=k}\|\phi_t\|^2_{(V^{\mathrm{ex}}_k)^{-1}}},$$

using monotonicity of $\beta_n$ in $n$. By Lemma 17 the inner square root is at most $\sqrt{2dL}$ with $L = \log(1 + T/(d\varrho))$. Since $\sum_k n_k(T) = T$, concavity of $\sqrt{\cdot}$ (Cauchy–Schwarz on the vector of $\sqrt{n_k}$) gives $\sum_k \sqrt{n_k(T)} \leq \sqrt{KT}$. Hence

$$\sum_t w^{\mathrm{ex}}_t \;\leq\; 2\beta_T\sqrt{2dKTL} \;=\; B_{\mathrm{ex}}(T). \tag{20}$$

*Step 3b: the pooled branch, and where nesting enters.* Here we use Assumption 4. After $m$ observations have entered level $k$'s pooled design, standard matrix concentration under 4 gives $\lambda_{\min}(V_k^{\mathrm{po}}) \geq \varrho + \sigma^2 m/2$ once $m \geq m_0 = O(\sigma^{-4}\log(dT/\delta))$; the first $m_0$ rounds per level contribute at most $Km_0$ to the regret, a $T$-independent additive term we absorb into the $\widetilde{O}$. Beyond that, $\|\phi_t\|_{(V_k^{\mathrm{po}})^{-1}} \leq \sqrt{2/(\sigma^2 m_k(t))}$. Now the key point. Level $k$'s pooled design receives a sample on *every* round in which any level $i \geq k$ is played, so with $n_i$ the final play counts, $m_k(T) = \sum_{i\geq k} n_i$, which is non-increasing in $k$ with $m_1(T) = T$ and $m_{K+1} := 0$. Grouping the sum by played level,

$$\sum_t 2\beta_m \|\phi_t\|_{(V_{k_t}^{\mathrm{po}})^{-1}} \;\leq\; 2\beta_T\sqrt{\frac{2}{\sigma^2}}\sum_{k=1}^{K}\frac{n_k}{\sqrt{m_k(T)}},$$

and since $n_k = m_k - m_{k+1}$ and $u \mapsto u^{-1/2}$ is decreasing,

$$\sum_{k=1}^{K}\frac{m_k - m_{k+1}}{\sqrt{m_k}} \;\leq\; \sum_{k=1}^{K}\int_{m_{k+1}}^{m_k}\frac{du}{\sqrt{u}} \;=\; \int_0^{m_1}\frac{du}{\sqrt{u}} \;=\; 2\sqrt{T}, \tag{21}$$

where the inequality holds termwise because on $[m_{k+1}, m_k]$ we have $u^{-1/2} \geq m_k^{-1/2}$, and the telescoping uses that the intervals $[m_{k+1}, m_k]$ partition $[0, m_1] = [0, T]$. This is exactly (16), and it is available *because* the feedback is lower-triangular: it is what makes $m_k$ decreasing with $m_1 = T$.

For the bias part of $w_t^{\mathrm{po}}$, Step 1 gives the per-round contribution $\delta_{\max}\sqrt{m_{k_t}}\,\|\phi_t\|_{(V_{k_t}^{\mathrm{po}})^{-1}}$. Past the burn-in, $\lambda_{\min}(V_k^{\mathrm{po}}) \geq \sigma^2 m_k/2$ gives $\|\phi_t\|_{(V_{k_t}^{\mathrm{po}})^{-1}} \leq \sqrt{2/(\sigma^2 m_{k_t})}$, so the $\sqrt{m_{k_t}}$ cancels and each round contributes at most $\delta_{\max}\sqrt{2}/\sigma$, independent of how much data has accumulated. This cancellation is the formal statement of the saturation described in Remark 19: additional free observations shrink the stochastic width but not the bias, so the pooled interval stops improving at width $\asymp \delta_{\max}/\sigma$. Summing over $T$ rounds gives $\delta_{\max}T\sqrt{2}/\sigma$. Collecting,

$$\sum_t w_t^{\mathrm{po}} \;\leq\; 2\beta_T\sqrt{\frac{2T}{\sigma^2}} + \delta_{\max}\,T\,\frac{\sqrt{2}}{\sigma} \;=\; B_{\mathrm{po}}(T). \tag{22}$$

*Step 4: combining, and the per-level form.* The two branches (20) and (22) are both valid, so the regret is bounded by twice their minimum, the factor 2 absorbing the $O(Km_0)$ burn-in and union-bound constants. This gives the scalar form $R_T \leq 2\min\{B_{\mathrm{ex}}, B_{\mathrm{po}}\}$.

For the per-level form (14), observe that the choice of which channel to admit is made *independently for each level*: the intersection (11) at level $k$ uses whichever channels are available for $k$, and the width bound (19) is applied per round to the played level. Fix a trust set $\mathcal{T}$ and run the algorithm admitting the pooled channel only for $k \in \mathcal{T}$. Partition the rounds by whether the played level lies in $\mathcal{T}$. Rounds playing $k \notin \mathcal{T}$ are bounded by Step 3a restricted to those levels, contributing $2\beta_T\sqrt{2dTL}\,|[K]\setminus\mathcal{T}|^{1/2}$ (the same argument with $\sum_{k\notin\mathcal{T}}\sqrt{n_k} \leq \sqrt{|[K]\setminus\mathcal{T}|\,T}$). Rounds playing $k \in \mathcal{T}$ are bounded by Step 3b applied to level $k$ alone: its stochastic part contributes $2\beta_T\sqrt{2T/\sigma^2}/\sqrt{K}$ after the telescoping (21) is distributed across the $K$ levels, and its bias part contributes $\delta_k T\sqrt{2}/(K\sigma)$ by the cancellation in Step 3b with $\delta_k$ in place of $\delta_{\max}$. Summing and minimizing over $\mathcal{T}$ gives (14); taking $\mathcal{T} = \emptyset$ and $\mathcal{T} = [K]$ recovers the two scalar branches.

Claim (a) is the $\mathcal{T} = \emptyset$ branch. Claim (b) follows by comparing level $k$'s two terms: its pooled term is smaller iff $\delta_k T\sqrt{2}/(K\sigma) \lesssim 2\beta_T\sqrt{2dTL}/\sqrt{K}$, i.e. $\delta_k \lesssim \sigma\beta_T\sqrt{dKL/T}$ up to constants. Claim (c) follows from Step 3b's per-round comparison: level $k$'s stochastic half-width is $\beta_{n_k}\sqrt{2/(\sigma^2 n_k)}$ and its bias half-width is the constant $\delta_k\sqrt{2}/\sigma$, so the former exceeds the latter iff $\beta_{n_k}^2/n_k \geq \delta_k^2$, i.e. iff $n_k \leq 2\beta_T^2/\delta_k^2 = n_k^\star$ (using $\beta_n \leq \beta_T$ and absorbing the factor 2). Past $n_k^\star$ the bias term dominates and the interval is pinned at $\delta_k\sqrt{2}/\sigma$ regardless of $n_k$. No step used Assumption 1. □

*Remark* 20 (Where the proof is loose). Two places. (i) In Step 1 the bias residual is bounded by $\delta_{\max}$ plus a term of the same order rather than exactly $\delta_{\max}$; we absorbed the difference into the leading constant 2, which is why (14) carries an explicit factor 2 rather than being stated with sharp constants. (ii) Step 3b's burn-in $m_0$ is not optimized. Neither affects the rates or the crossover location.

### L.5 Proof of Theorem 12

*Proof.* Soundness ($k^\star_\lambda(x_t) \in \mathcal{W}_t$ always, on $\mathcal{E}$) was established in Step 2 of the previous proof: since $\mathrm{LCB}_j - \lambda c_j \le u^\star_\lambda(x_t)$ for all $j$ and $u^\star_\lambda(x_t) \le q(x_t, k^\star) - \lambda c_{k^\star} \le 1 - \lambda c_{k^\star}$, the cost-domination test never excludes $k^\star$. Hence pruning can only remove suboptimal levels and the regret decomposition (19) is unaffected.

For the count, fix a level $k \notin \mathcal{W}(x)$, so $1 - \lambda c_k < u^\star_\lambda(x)$, and let $\Delta_k(x) = u^\star_\lambda(x) - u_\lambda(x,k) \ge u^\star_\lambda(x) - (1 - \lambda c_k) > 0$. Playing $k$ at round $t$ requires, by the selection rule and (19), that $w_t \ge \Delta_k(x_t) \ge \Delta_{\min}$. On $\mathcal{E}$, $w_t \le 2\beta_T \|\phi_t\|_{(V_k^{\mathrm{ex}})^{-1}}$, so each such round has $\|\phi_t\|^2_{(V_k^{\mathrm{ex}})^{-1}} \ge \Delta^2_{\min}/(4\beta_T^2)$. Summing over those rounds and applying Lemma 17 to level $k$'s native design,

$$N_k \cdot \frac{\Delta^2_{\min}}{4\beta_T^2} \;\le\; \sum_{t:k_t=k} \|\phi_t\|^2_{(V_k^{\mathrm{ex}})^{-1}} \;\le\; 2d\log\bigl(1 + T/(d\varrho)\bigr),$$

so $N_k \le 8\beta_T^2 d\log(1 + T/(d\varrho))/\Delta^2_{\min}$. Summing over the at most $K$ such levels gives the stated $O(\beta_T^2 d \log T/\Delta^2_{\min})$ total (with the $K$ absorbed, as each level's count is bounded individually and $K$ is a constant of the instance). Since all but $O(\beta_T^2 d\log T/\Delta^2_{\min})$ rounds play a level in $\mathcal{W}(x_t)$, the grouping in Step 3a may be restricted to those levels, replacing $\sum_k \sqrt{n_k} \le \sqrt{KT}$ by $\sqrt{WT}$.

Finally, the test itself uses only $c_k$ (known), the range $q \le 1$, and $\mathrm{LCB}_j$ for other levels: no observation of $k$ and no monotonicity assumption. □

### L.6 Proof of Proposition 13

*Proof.* Fix any level $j$ and any $k > j$. Under $\mathcal{I}_A$ the truncated observation of level $j$ has mean $v$ (quality $v$, zero bias). Under $\mathcal{I}_B$ it has mean $(v + \delta_j) - \delta_j = v$. Both instances satisfy Assumption 2 with the same bound $\delta_j$, since $|-\delta_j| \le \delta_j$; note this uses only the two-sided form of (5), not any sign restriction. If, as in our verifiable-reward setting, the score is Bernoulli, both are $\mathrm{Bern}(v)$, so the two induced laws are equal and their total variation distance is 0. Since the truncated observations at distinct levels are conditionally independent given $x_t$ and $k_t$ in both instances, the *joint* law of the whole truncated vector $(\widetilde{Y}_t(j))_{j<k}$ is the product $\mathrm{Bern}(v)^{\otimes(k-1)}$ under both, again with total variation distance 0. (Our verification computes this joint distance by enumerating all $2^{K-1}$ outcomes and confirms it is exactly 0.)

Therefore any algorithm whose level choices are measurable functions of truncated observations, contexts and costs alone induces identical distributions over play sequences on $\mathcal{I}_A$ and $\mathcal{I}_B$. Now choose costs so that the instances disagree about the optimum. Let $K = 2$, $j = 1$, $k = 2$, and set $\lambda c_2 - \lambda c_1 = \delta_j/2$ with $q(\cdot, 2)$ equal in both instances to $v + \delta_j/2$. Then under $\mathcal{I}_A$,

$$u_A(1) = v - \lambda c_1, \qquad u_A(2) = \bigl(v + \tfrac{\delta_j}{2}\bigr) - \lambda c_2 = v - \lambda c_1,$$

a tie, so perturbing $v$ by an arbitrarily small amount makes level 2 strictly optimal under $\mathcal{I}_A$. Under $\mathcal{I}_B$, where level 1's true quality is $v + \delta_j$,

$$u_B(1) = \bigl(v + \delta_j\bigr) - \lambda c_1, \qquad u_B(2) = v - \lambda c_1, \qquad u_B(1) - u_B(2) = \delta_j,$$

so level 1 is strictly optimal under $\mathcal{I}_B$ by margin $\delta_j$. Because the algorithm cannot distinguish the instances, on at least one of them it plays the locally suboptimal level with probability at least $1/2$ on every round, incurring expected per-round regret at least $\delta_j/2$ and hence $R_T = \Omega(\delta_j T)$.

Adding native observations breaks the tie: under $\mathcal{I}_A$ a native pull of level $j$ has mean $v$, under $\mathcal{I}_B$ it has mean $v + \delta_j$, and these are distinguishable in $O(\delta_j^{-2}\log(1/\delta))$ pulls. This is precisely why TruncUCB retains the native channel, and why the $\delta_j$ terms in (14) cannot be removed by any amount of free feedback — the formal counterpart of the saturation threshold $n_j^\star = 2\beta_T^2/\delta_j^2$ in (15). □

### L.7 Proof of Proposition 14

*Proof.* Take the stated instance: $K = 3$, $q = (0.90, 0.35, 0.60)$, $c = (0.90, 0.95, 1.00)$, $\lambda = 0.1$, hence $u = (0.81, 0.255, 0.50)$ and $k^\star = 1$ uniquely, with gap $\Delta = u_1 - u_3 = 0.31$. Note that no

level is cost-dominated here: $1 - \lambda c_k = (0.91, 0.905, 0.90)$ all exceed $u^\star = 0.81$, so Theorem 12 prunes nothing and the monotone rule is what governs behaviour. Assumption 1 is violated by $q_1 - q_2 = 0.55$.

Consider the rule $\mathrm{U}_j \leftarrow \min\{\mathrm{U}_j, \mathrm{U}_{j+1} + \varepsilon\}$ applied downward for $j = K-1, \ldots, 1$, with $\varepsilon$ calibrated from a measured aggregate violation rate (in our pilot $\varepsilon = 0.046$). Suppose levels 2 and 3 have been sampled enough that their intervals have width $\leq \varepsilon$, so $\mathrm{U}_3 \leq q_3 + \varepsilon = 0.60 + \varepsilon$ and $\mathrm{U}_2 \leq q_2 + \varepsilon$. The rule then caps

$$\mathrm{U}_2 \leftarrow \min\{\mathrm{U}_2,\ \mathrm{U}_3 + \varepsilon\} \leq 0.60 + 2\varepsilon, \qquad \mathrm{U}_1 \leftarrow \min\{\mathrm{U}_1,\ \mathrm{U}_2 + \varepsilon\} \leq 0.60 + 3\varepsilon.$$

Take $\varepsilon = 0.05$, the order of magnitude a practitioner would calibrate from a measured violation rate (ours are $8.6\%$ and $17.5\%$).

*Step 1: the cap destroys validity at the optimal level.* Once levels 2 and 3 have been sampled enough that their own intervals are tight, $\mathrm{U}_2 \approx q_2 = 0.35$ and $\mathrm{U}_3 \approx q_3 = 0.60$. The downward cap then forces

$$\mathrm{U}_2 \leftarrow \min\{\mathrm{U}_2, \mathrm{U}_3 + \varepsilon\}, \qquad \mathrm{U}_1 \leftarrow \min\{\mathrm{U}_1, \mathrm{U}_2 + \varepsilon\} \ \approx\ q_2 + \varepsilon \ =\ 0.40,$$

because the binding term for $\mathrm{U}_1$ is the *intermediate* level's tight upper bound $q_2 = 0.35$, not level 3's. Since $q_1 = 0.90 \gg 0.40$, the interval for level 1 excludes its own true quality. The cap is recomputed every round from levels $2, 3$ alone, so no amount of sampling of level 1 can repair it. This is already a failure of the confidence construction: the algorithm holds $\mathrm{U}_1 < \mathrm{L}_1$, an empty interval.

*Step 2: the ranking inverts.* With $\mathrm{U}_1 \approx q_2 + \varepsilon$, level 1's optimistic index is $\mathrm{U}_1 - \lambda c_1 \approx 0.40 - 0.09 = 0.31$, while level 3's index converges to $u_3 = 0.60 - 0.10 = 0.50 > 0.31$. Level 3 is therefore selected whenever both are live, and level 1 — the unique optimum, $u_1 = 0.81$ — is starved. Each such round incurs instantaneous regret $\Delta = u_1 - u_3 = 0.31$.

*Step 3: the starvation is self-sustaining.* Level 1 receives no new native samples once it stops being selected, so $\mathrm{U}_1$ cannot rise; levels $2, 3$ continue to be sampled, so the cap does not loosen. Hence the inverted ranking persists for all subsequent rounds and $R_T = \Omega(T)$ with rate $\Delta = 0.31$ per round.

Our simulation confirms every step. At $T = 6 \times 10^4$ on this instance the capped variant holds, at a representative context, $\mathrm{U}_1 = 0.578$ against $\mathrm{L}_1 = 0.784$ — the empty interval of Step 1 — with optimistic indices $(0.488, 0.350, 0.574)$, so level 3's index exceeds level 1's exactly as Step 2 predicts. Averaged over seeds it plays level 3 on $65.1\%$ of rounds and level 1 on $34.5\%$, with doubling ratio $R(T)/R(T/2) = 3.33$ and $22.4\times$ the regret of the uncapped default ($1.21 \times 10^4$ versus $540$; the default's ratio is $1.03$, and the GUARDED mode matches the default exactly). We note that this failure requires the per-level bias profile we measured: with a single uniform bias across levels the capped rule's regret is indistinguishable from the default's, so the counterexample is calibrated to the measurement rather than to a worst case chosen for effect.

For TRUNCUCB in mode OFF, no such cap exists and Theorem 8 applies verbatim. In mode GUARDED, only *lower* bounds propagate *upward* ($\mathrm{L}_k \leftarrow \max\{\mathrm{L}_k, \mathrm{L}_{k-1} - \varepsilon\}$), which can only raise a lower bound; it never caps the optimal level's optimism, so the validity argument of Step 1 of Theorem 8's proof is untouched and the bound continues to hold.

□