Should You Pay for More Thinking? Low-Budget Probes Are Reliable Purchasing Signals but Poor Forecasts

Tony Salomone
Transformer Lab
tony.salomone@lab.cloud

Deep Gandhi
Transformer Lab

Ali Asaria
Transformer Lab

Abstract

Evaluating reasoning models at large per-response thinking budgets is expensive, so practitioners would like cheap low-budget evaluations to stand in for them. We show that a probe restricted to at most 4k thinking tokens per response can reliably answer the practitioner's operative question, whether paying for a higher budget will improve accuracy, even though it cannot predict the resulting score. The probe itself reveals which regime a model and benchmark occupy: when the model does not consume the probe's thinking allowance, higher budgets delivered at most small gains (the largest was 0.051 in the preregistered suite, 0.052 across the completed campaign), and when it saturates the allowance, measured high-budget accuracy met or exceeded the frozen curve forecast in all 14 such cells of the preregistered suite. The resulting decision rule (buy more budget only when the probe saturates its cap and the frozen curve forecasts a gain of at least δ) makes zero over-promise errors across its 14 buy decisions at $\delta = 0.02$ and its 12 at $\delta = 0.05$ (rule-of-three 95% upper bounds 21% and 25%). The rule was constructed on the 24 Qwen3 cells; the held-out R1-Distill lineage has since completed, delivering six holdout buy decisions, all correct: one measurement landed within 0.0005 of its frozen forecast, and on one buy cell the measured score fell 0.060 below its frozen forecast while the buy remained correct by a wide margin, the only substantive forecast shortfall on a buy cell in the campaign. As a numeric forecaster, however, the probe fails its preregistered bars: the best frozen predictor lands within twice seed noise on only 35.7% of budget-hungry cells, frozen 90% intervals cover 53.6% of 28 outcomes against an 80–97% target, the misses are systematically one-sided on budget-hungry cells (cluster-level sign test $p = 0.031$), and probe cost misses its 20% token bar (35–39% of the high-budget suite; about 21% in GPU-hours). All predictions were fit on the realized thinking tokens of low-budget cells and frozen and hashed before any high-budget cell for that model lineage ran, across three open models and four reasoning benchmarks under a deployable hard thinking-budget cap. In our suite, low-budget probes are conservative purchasing signals rather than forecasters.

1 Introduction

Test-time compute has become a primary scaling axis for reasoning models [1]: given more per-response thinking tokens, models such as Qwen3 [2] and DeepSeek-R1 [3] solve substantially harder problems. The same trend makes evaluation expensive. A single 32k-token-budget sweep of one 14B model over four standard reasoning benchmarks with seed replication costs tens of GPU-hours; a full budget grid for several models runs to a hundred or more. An evaluation practitioner deciding

whether to pay for a high-budget run therefore faces a natural question: *can a cheap probe at small thinking budgets predict what the model would score at a much larger budget, so that the expensive run need not happen at all?*

Prior work supplies the ingredients but not the answer. Log-linear accuracy-in-log-budget curves are documented on our model families [4, 5, 6] and derived theoretically [7], but the published fits interpolate within measured grids rather than validating extrapolation to held-out budgets. Extrapolation with held-out validation exists on neighboring axes, parallel sampling [8] and training compute [9, 10, 11], but, to our knowledge, not on the sequential thinking-budget axis, a central framing of test-time compute [12]. Whether the underlying curves are even monotone is contested [13, 14, 15], though much of the non-monotone evidence conditions on endogenous realized length rather than an exogenous budget knob [16].

We run the missing experiment, and we run it in a form that leaves no room for post-hoc adjustment. For each of three models (Qwen3-8B, Qwen3-14B, DeepSeek-R1-Distill-Qwen-14B) we measure a low-budget probe grid (budgets 0 to 4096 thinking tokens) on four benchmarks (AIME 2024 and AIME 2025 [17, 18], MATH-500 [19, 20], GPQA-Diamond [21]), fit a shortlist of curve families on the realized thinking tokens of the probe cells, select a family by leave-one-cell-out error, and freeze the resulting predictions: SHA-256 hashes are recorded before any cell above 4096 tokens for that model is queued, and the artifact is committed to a version-controlled repository before any of that model's high-budget measurements existed. Only then do we measure the held-out cells at 8192, 16384, and 32768 tokens.

The result is a two-part finding, and both parts are the contribution.

The forecast fails. Against the preregistered gate (prediction within twice the target cell's seed-to-seed standard deviation), the best frozen approach passes on about a third of budget-hungry cells, frozen 90% intervals cover roughly half of the 28 held-out outcomes, and the probe misses its preregistered token-denominated cost bar (exact figures in the abstract and Section 5.2). The misses are systematic, with three diagnosed mechanisms: curve-shape blindness on budget-hungry benchmarks (measured exceeds the frozen forecast in all six independent model–benchmark curve groups, with under-prediction up to +0.24), axis mismatch on budget-indifferent cells (models stop consuming budget near 4–6.6k realized tokens, so nominal-budget extrapolation over-predicts, while the same frozen curve read post hoc at the realized-token axis is consistent with the measurement on every such Qwen3 cell), and a mild, saturating overthinking penalty (MATH-500 declines by at most 0.023 from its 4k peak and partially recovers).

The purchasing signal works, within its stated exposure. A one-line saturation check computable from the probe alone (does mean realized thinking at the 4k probe reach 90% of the cap?) separates budget-hungry from budget-indifferent cells and turns the probe into a decision rule: buy more budget only if the probe is saturating and the frozen curve forecasts a gain of at least δ. Across the 28 validated cells the rule makes zero over-promise errors in its 14 buy decisions at the lower threshold and its 12 buy decisions at the higher one, where the naive curve-only rule over-promises 7 and 5 times; every residual error of the saturation-aware rule is a cheap missed upside. Zero failures in 12–14 clustered decisions bounds the true over-promise rate only loosely, so Section 5.7 reports exposure sets, rule-of-three bounds, and interval estimates rather than bare proportions. On budget-hungry cells of the preregistered suite the probe forecast acted as a conservative lower bound: it never over-sold the gain there. The completed holdout adds one exception, a buy cell whose measured score fell 0.060 below its frozen forecast while the buy remained correct by a wide margin (Section 5.7). The screen and rule were constructed after the 24 Qwen3 cells had been

measured, and were fixed roughly three hours before the first held-out measurement of the third model lineage completed; that lineage is a genuine holdout, and all six of its buy decisions landed correct.

Our contributions, each falsifiable against the recorded artifacts:

1. A preregistered freeze-then-measure protocol for thinking-budget extrapolation: frozen, hashed predictions from probe-only fits, validated on 28 held-out high-budget cells across three model lineages (Sections 3, 5.2).

2. A characterization result: under an exogenous budget knob, accuracy-versus-realized-thinking curves are smooth and monotone-increasing endpoint-to-endpoint on all 12 model–benchmark pairs, up to within-noise local dips, a small underthinking-cliff region at $B = 512$, and one mild, bounded MATH-500 overthinking dip, with harness-enforced cap compliance (Control $= 1.0$) on every measured cell[1] and no truncation paradox at the smoke gate (Section 5.1).

3. An honest negative: the preregistered point-prediction and calibration gates fail, with a mechanism-level account of why (curve-shape blindness, axis mismatch, mild overthinking) and a moderator map over budget gap and benchmark class (Sections 5.2–5.5).

4. A positive reframing with explicit exposure accounting: a probe-computable saturation screen (2.8× error separation on the preregistered noise-normalized axis; the raw-error separation reverses, Section 5.2) and a buy rule with zero over-promises in 14 and 12 buy decisions at the two tested thresholds. The rule was constructed on the 24 Qwen3 cells and is held out on the R1-Distill lineage, where the completed holdout delivered six buy decisions, all correct (Section 5.7).

5. An approach comparison under preregistration: the aggregate selected-family curve beats a per-item logistic alternative (16/24 versus 10/24 CI coverage at the post-hoc realized-token diagnostic readout), which wins exactly the one cell the preregistration flagged in advance (Section 5.6).

2 Related Work

Test-time scaling and thinking budgets. Scaling test-time compute is established as an alternative to scaling parameters [1], with a survey cataloging the mechanisms for controlling reasoning effort [12] and a broad empirical map of reasoning-model behavior across models and benchmarks [22]. On the sequential axis specifically, budget forcing [23], RL-trained length control [4], and adaptive effort control [24] all modulate thinking length, while prompt-difficulty routing modulates inference cost through model choice rather than thinking length [25]; log-linear accuracy gains in log-budget are reported on medical reasoning [6], and log-linear pass@k gains on fractured chain-of-thought sweeps [5]. A theoretical account derives the log-linear regime from difficulty mixtures [7]. None of these works fits on cheap low-budget cells and validates extrapolation on held-out high-budget cells; that validation gap is the specific opening this paper addresses.

[1]The 96 grid cells with Control $= 1.0$ decompose as 60 low-grid cells (3 models × 5 budgets × 4 benchmarks) plus all 36 measured high-budget cells (the 28-cell preregistered suite and the 8 R1-Distill holdout cells); the 6 smoke cells (2 lineages × 3 budgets) also passed.

The curve-shape debate. Whether more thinking monotonically helps is contested. Inverted-U relationships with difficulty-dependent peaks [13], flat or oscillatory budget curves on o1-like models [14], declining accuracy in realized length within difficulty tiers [16], model-family splits between monotone and peaked behavior [15], collapse regimes at high complexity [26], flat curves on knowledge-intensive tasks [27], and plateau-then-jump shapes [28] all argue against naive extrapolation. A key reconciliation, which our design operationalizes, is that much of the non-monotone evidence conditions on endogenous realized length (models think longer on items they cannot solve), not on an exogenous budget intervention [16]. Our exogenous-knob sweep de-confounds this, and our finding of smooth, monotone-saturating curves (up to a mild MATH-500 dip) under an exogenous knob (Section 5.1) supports the artifact interpretation of much of the pathology literature, consistent with evidence that truncation-style knobs sit far below native control at small budgets and can act as artifactual performance ceilings [4, 29] and that incomplete reasoning can score below no reasoning at all [30].

Performance prediction and scaling-law extrapolation. Fit-cheap-extrapolate-expensive methodology is mature on other axes: broken neural scaling laws [31], Bayesian extrapolation with prior-data fitted networks [32], latent-variable scaling frameworks [33], budget-efficient fit-point selection [9], prescriptive scaling with evaluation-budget-optimal designs [10], and observational scaling laws across model families [11]. On the parallel-sampling axis, per-item distributional modeling predicts pass@k scaling at large k from small-k measurements with up to 10× error reductions over aggregate regression [8]. We import this literature's discipline (held-out-budget validation, uncertainty-first fitting, preregistration of frozen fits) onto the sequential thinking-token axis, where it has not been applied, and we test the per-item recipe head-to-head against aggregate curve fitting (Section 5.6).

Efficient evaluation and its failure modes. Reduced-cost evaluation via item subsets is well studied [34], and its central caution transfers to our setting: learned predictors that dominate in-distribution can invert to worst-in-class in extrapolation regimes, so estimators anchored on unbiased measurements degrade more gracefully [35]. Signal-to-noise analyses formalize when benchmark comparisons are decidable at all [36], and metric-induced [37] and seed-sampling-induced [38] artifacts can manufacture apparent discontinuities. These works motivate our noise-normalized success gate, our carry-forward baselines, and our screen-based reporting. Kinetics-style cost accounting, in which attention makes true cost quadratic rather than linear in tokens, informs our cost-ratio reporting [39]. Contamination concerns, acute for small competition benchmarks and nearly undetectable post hoc in reasoning models [40], motivate reporting AIME 2024 and AIME 2025 as separate strata. Hybrid think/no-think switching is imperfect, so the zero-budget anchor is treated as diagnostic only [41]. Broader arguments for forecastable evaluation practice appear in Anwar et al. [42].

3 Method

3.1 Budget knob and knob-fidelity gate

The controlled variable is the nominal per-response thinking budget B: a hard cap on thinking tokens enforced in a two-phase serving harness. Phase one generates the thinking segment under a token cap with a forced transition to answering when the cap is exhausted; this mechanism is

identical for both lineages at positive budgets, and the lineages differ only at the $B = 0$ anchor (Qwen3's native no-think mode versus a template-forced empty thinking segment for R1-Distill). Phase two generates the visible answer. Because both lineages are evaluated under the same hard cap with forced transition, every curve in this paper measures the *cap-budget response*: accuracy as a function of an enforced serving-side budget, which is the control a deployment can actually turn. We make no claim about the response to natively trained budget controls; at small caps the probe cells partly score amputated chains of thought, and the planned native-versus-truncation ablation that would quantify the difference between the two constructs was not run (Section 7). The measured covariate is the *realized* mean thinking tokens $\bar{x}$ of each cell, which can fall well below B when the model stops thinking early. Following repeated warnings in the literature against fitting on nominal budget [4, 39, 26], all curve fits regress accuracy on realized tokens; the nominal-to-realized mapping is reported separately.

Before any sweep, each knob lineage passes a smoke gate: (i) *Control*, the fraction of responses whose realized thinking respects the budget, must be at least 0.90 at positive budgets; (ii) no truncation paradox: accuracy at the smoke's smallest positive budget (1024) must not fall below the $B = 0$ anchor [30]; (iii) the top grid cell must truncate fewer than 10% of responses, certifying it unconstrained [29]; this third criterion is checkable only once the top cell lands. Qwen3-14B was not smoked separately; its low grid was queued on the strength of the Qwen3-8B lineage smoke. Criteria (i) and (ii) passed for both smoked lineages at smoke time; criterion (iii) passed post hoc on all three lineages once each top cell landed (Section 5.1).

3.2 Budget grid and probe/holdout split

The nominal grid is $B \in \{0, 512, 1024, 2048, 4096, 8192, 16384, 32768\}$, log-spaced. Cells with $B \leq 4096$ form the probe; cells at $B \in \{8192, 16384, 32768\}$ are the held-out prediction targets. The $B = 0$ no-think cell is a diagnostic anchor and is excluded from all fits [41]. The unit of prediction is benchmark-level accuracy at budget B. A (model, benchmark, budget) cell is entirely probe or entirely holdout; no quantity derived from any high-budget measurement enters any fit, enforced by run order (Section 3.4).

3.3 Curve families and selection

For each (model, benchmark) pair, four candidate families are fit to the probe cells (512–4096) at realized-token x: (a) linear in $\log x$; (b) floored sigmoid in $\log x$ with the guessing floor fixed rather than fitted (0.25 for GPQA multiple choice, 0 for AIME/MATH) [33, 11]; (c) a saturating power law on error rate; (d) a broken-scaling form capable of expressing breaks [31], acknowledged near-degenerate on four probe points. Fitting uses multi-start bounded L-BFGS; family selection minimizes leave-one-cell-out (LOO) mean squared error within the probe range, with ties resolved toward fewer parameters. Selection never sees held-out budgets. Uncertainty is propagated by an item-level bootstrap (200 replicates) through the selected family, yielding a frozen 90% prediction interval per target cell.

Two alternatives are preregistered alongside this primary approach (A): (B) a per-item shared-slope logistic in $\log x$, the sequential-axis analogue of the distributional recipe that dominates on the parallel axis [8], aggregated to benchmark accuracy as a mixture; and (C) an anchored convex blend of the carry-forward baseline and the selected curve, weighted by inverse LOO error, following the graceful-degradation template of Polo et al. [34] and Zhang et al. [35]. The plan mandated three

naive baselines: last-observed-value carry-forward, ordinary least squares linear in log realized tokens (coinciding with family (a), whose per-cell fits and frozen predictions are recorded in the preregistration artifacts), and a 2×-probe extrapolation. The third was never computed before freezing; Section 5.2 reports it as a post-hoc comparison computed from the landed 8k cells.

Readouts. Every frozen curve is a function of realized thinking tokens x, and it is evaluated in two distinct ways that the reader should keep separate. The *protocol forecast* Pred@B evaluates the frozen curve at $x = B$, the nominal budget of the target cell; it is available at probe time, before the target cell exists, and it is the object the preregistered gate scores. The *diagnostic readout* Pred@$\bar{x}$ evaluates the same frozen curve at $\bar{x}$, the target cell's measured mean realized thinking tokens. Because $\bar{x}$ is a statistic of the held-out run itself, Pred@$\bar{x}$ is not available at probe time and is never used as a forecast; it serves only post hoc, to attribute a miss to the axis (the nominal-to-realized mapping) versus the curve shape. All headline verdicts score Pred@B.

3.4 Preregistration protocol

For each model lineage, the full prediction artifact (all candidate-family parameters and LOO errors, selected-family point predictions at 8192/16384/32768, 90% bootstrap intervals, per-item and anchored alternates, carry-forward values, and a probe-cost sweep over $B_{\text{low}} \in \{1024, 2048, 4096\}$) was serialized, hashed with SHA-256, and recorded *before any cell above 4096 tokens for that lineage was queued*; the commit of each artifact to a version-controlled repository landed before any high-budget measurement for that lineage existed (for the two Qwen3 lineages, the commit also preceded queuing). At evaluation time, the approach-A curves were reloaded directly from the hashed artifacts without refitting, and refitting the frozen per-item code on the frozen probe data reproduced every committed per-item prediction with zero mismatches. Headline results use $B_{\text{low}} = 4096$. The frozen artifacts also recorded observables to watch: the R1-Distill freeze flagged its own MATH-500 log-linear extrapolation (0.955 at 32k) as implausibly high and named the per-item alternative (approximately 0.79) the believable forecast, an A-versus-B observable resolved in Section 5.6.

3.5 Saturation screen and decision rule

The saturation screen asks whether extrapolation should be trusted at all, using probe data only; it was formulated after the 24 Qwen3 high-budget cells had been measured, so it is preregistered only with respect to the R1-Distill lineage. The screen costs nothing: it reads a statistic the probe already logs. Stated with probe-time inputs only, the full rule is:

1. Run the low-budget probe grid ($B \in \{512, 1024, 2048, 4096\}$ plus the $B = 0$ anchor) with seed replication, logging realized thinking tokens per response.
2. Compute $\bar{x}_{4096}$, the mean realized thinking tokens of the $B = 4096$ probe cell, and a_{4096}, that cell's measured accuracy.
3. If $\bar{x}_{4096} < 0.9 \times 4096$, classify the cell *budget-indifferent* and predict no gain (do not buy).
4. Otherwise the cell is *budget-hungry*: fit the candidate families of Section 3.3 on the probe cells' realized tokens, select by LOO error, and compute the protocol forecast Pred@B at the target budget B. Buy if and only if Pred@$B - a_{4096} \geq \delta$.

Here $B \gg 4096$ is the candidate high budget, δ is the practitioner's minimum accuracy gain worth paying for, and a_{4096} is the *measured* probe-top accuracy, not the fitted curve's value there. Outcomes are scored against the measured gain $a_B - a_{4096}$, where a_B is the point estimate of measured accuracy at the target cell. Errors divide into over-promises (rule says buy, measured gain below δ: the costly error, paying for nothing) and under-promises (rule says no buy, measured gain at least δ: a missed upside). The rule was fixed in the analysis record at 00:15 UTC, roughly three hours before the first R1-Distill high-budget cell completed at 03:13 UTC, so the third lineage is a genuine out-of-sample test of the rule.

3.6 Success criteria

Seven criteria were preregistered. The gating bar is noise-normalized: the protocol forecast Pred@B must land within twice the target cell's seed-to-seed standard deviation σ_{seed}, for the majority of well-behaved cells (Section 5.2 additionally reports a non-preregistered pooled-σ reading of the same gate, because the per-cell σ_{seed} rests on as few as 4 seeds). Absolute practitioner bars (± 5 points AIME, ± 3 points MATH-500/GPQA) are reported alongside but do not gate, because a fixed point bar on a 30-item benchmark partly measures the seed budget rather than the method. Further criteria: screen conditional purity of at least 90%; stratified reporting that excludes flat cells from the headline; 90% interval coverage in 80–97%; probe cost at most 20% of the full evaluation; the best method must beat every computed naive baseline on mean noise-normalized error (Section 3.3 discloses the deferred 2×-probe baseline); and screen-flagged cells must show at least twice the extrapolation error of screen-passing cells, with the error measured on the same noise-normalized axis as the gate.

4 Experimental Setup

Models. Qwen3-8B and Qwen3-14B [2] and DeepSeek-R1-Distill-Qwen-14B [3]: two knob lineages and three models, all served under the shared cap harness of Section 3.1, with the lineages differing only at the $B = 0$ anchor. Models were chosen for the knob, and 8B is the floor because smaller models plateau by roughly 5k thinking tokens [39]. A planned Qwen3-32B extension was cut on infrastructure grounds (Section 7).

Benchmarks. AIME 2024 (30 items) [17] and AIME 2025 (30 items) [18], reported as separate strata as a contamination diagnostic [40]; MATH-500 (500 items, difficulty levels 1–5) [19, 20]; and GPQA-Diamond (198 items) [21]: 758 items total, four benchmarks drawn from three widely used reasoning benchmark families (the two AIME years form one family, reported separately), spanning headroom-hard to deliberately saturated regimes. Raw files are SHA-256 hashed; prompts are frozen once in deterministic manifests and reused identically at every budget cell; GPQA options are shuffled once per item with an item-seeded generator and held fixed across all cells and models.

Serving and scoring. All generation uses vLLM 0.10.1 [43] (with transformers 4.56.2 [44]) on single NVIDIA A100 80GB instances, temperature 0.6, top-p 0.95, `max_model_len` 40960, and an answer phase capped at 2048 tokens; this is the sampling configuration recommended by the model providers and used by the cited evaluations [40, 5]. Seeds are fixed per repeat index: 16 seeds per cell for each AIME year, 8 for GPQA-Diamond, 4 for MATH-500. Every response logs its realized thinking tokens. Math answers are scored by boxed-answer extraction with math-verify 0.7.0 [45]

(version-pinned after an unpinned run silently fell back to exact matching and depressed MATH-500 scores by about 0.06); GPQA is scored by letter extraction. No LLM judge is used anywhere, to keep small budget-cell deltas free of judge noise [12].

Runs and cost. Each low-grid job produces 22,720 scored responses (4 benchmarks × 5 budgets × per-benchmark seeds); each high-budget cell job produces 4,544. Record files are SHA-256 hashed. The 8B probe cost 6.76 A100-hours against 31.5 A100-hours for its three high-budget cells. The preregistered evaluation suite, fixed at the analysis freeze, comprises 28 cells (12 per Qwen3 model, 4 for R1-Distill at 8k); the remaining eight R1-Distill cells (16k and 32k) were preregistered, completed after the freeze, and are reported as the completed holdout (Section 5.7). The full campaign measured 36 high-budget cells.

5 Results

5.1 Knob fidelity and curve characterization

The budget knob is faithful on both lineages: Control = 1.0 on all 96 grid cells, comprising the 60 low-grid cells (3 models × 5 budgets × 4 benchmarks) and all 36 measured high-budget cells (the 28-cell preregistered suite plus the 8 R1-Distill holdout cells), as well as on the 6 smoke cells; since the cap is harness-enforced, this certifies the implementation rather than any model-side budget-following behavior. No truncation paradox appeared at the preregistered smoke gate (deltas at 1024 versus 0: +0.033 for Qwen3-8B and +0.10 for R1-Distill; smoke accuracies rose 0.200/0.233/0.383 and 0.133/0.233/0.450 at budgets 0/1k/4k respectively). Top-cell truncation rates at 32k are below the 10% certification threshold on all three lineages (0.001–0.098 on the Qwen3 cells; 0.001–0.048 on the R1-Distill cells once its top cell landed), completing criterion (iii) everywhere.

Under this exogenous knob every probe curve is monotone-increasing endpoint-to-endpoint (slope signs positive on all 12 model–benchmark pairs, from +0.254 on R1 AIME-2024 down to +0.028 on Qwen3-14B MATH-500), with small mixed-sign deltas at $B = 512$ in the underthinking-cliff region, as the literature predicts [30]: at that budget, below the smoke gate's smallest cell, accuracy sits marginally below the $B = 0$ anchor in 6 of 12 model–benchmark pairs (deltas down to −0.027). A few further dips within seed-level noise appear along the curves (Qwen3-8B GPQA and Qwen3-14B AIME-2024 between 512 and 1024, and measured Qwen3-8B GPQA between 16k and 32k). The largest non-monotonicity in the campaign is the mild MATH-500 overthinking dip described in Section 5.4. This supports the hypothesis that reported non-monotone budget curves are largely artifacts of endogenous-length conditioning or forced truncation rather than properties of exogenous budget scaling [16, 4].

The realized-token distributions also justify the fit axis: at the 4096 probe cap, Qwen3-8B consumes essentially the full cap on AIME (about 4080 realized tokens) but only about 2936 on MATH-500, and R1-Distill only about 2225 on MATH-500. LOO family selection returned the same winners on all three lineages: floored sigmoid on both AIME years, log-linear on GPQA-Diamond and MATH-500.

5.2 The preregistered point-prediction gate fails

Table 2 reports the seven preregistered criteria; Table 1 gives the full 28-cell suite; Figure 1 shows the curves. The headline stratum is the 14 budget-hungry cells (the screen of Section 3.5 splits the

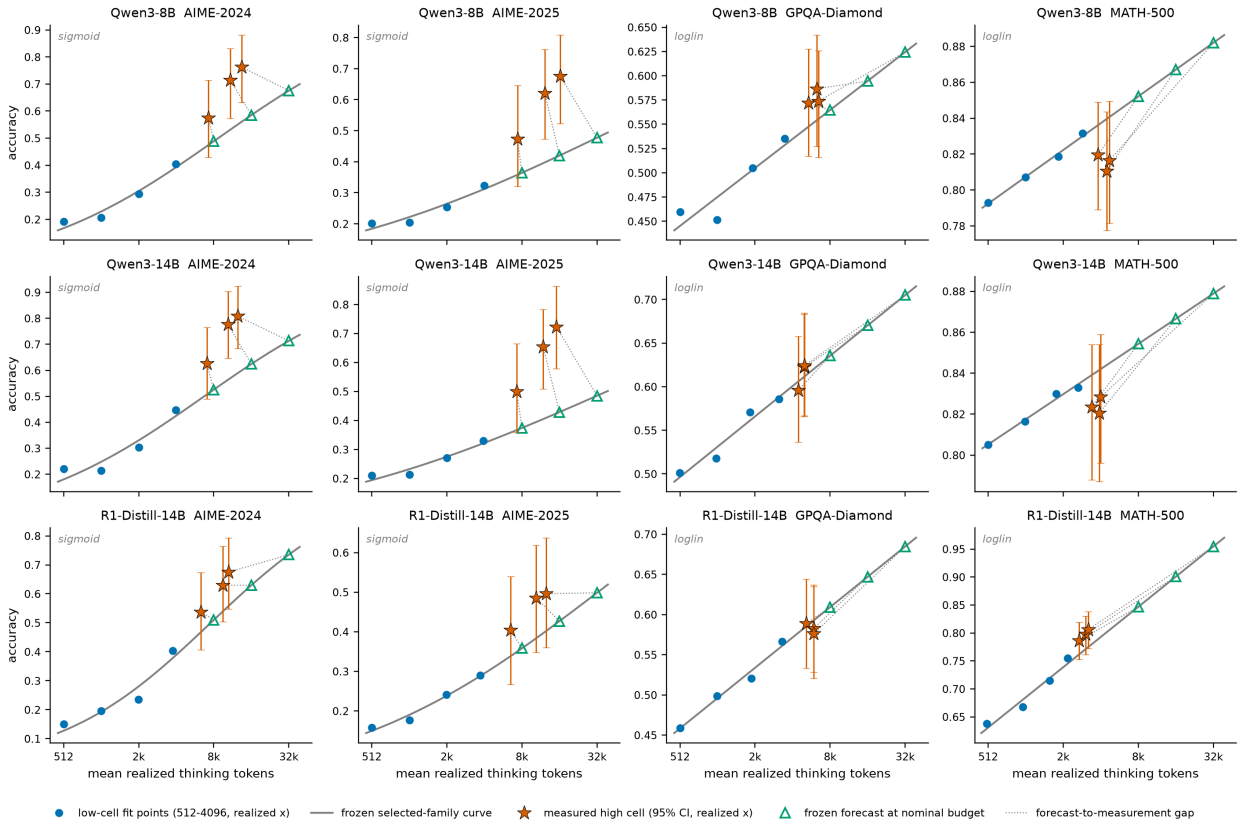

Figure 1: Accuracy versus realized thinking tokens for every model–benchmark pair: probe measurements, the frozen selected-family curve, and the held-out high-budget measurements. Budget-hungry benchmarks (AIME) keep climbing above the frozen curve until natural thinking saturates near 12–17k tokens; budget-indifferent benchmarks (MATH-500, Qwen3 GPQA) stop consuming budget near 4–6.6k realized tokens, where the frozen curve is accurate at the diagnostic realized-x readout and the miss is purely the axis.

suite 14/14). The best approach on that stratum, the protocol forecast Pred@B from the selected family, achieves mean noise-normalized error 2.67σ and passes the 2σ gate on 35.7% of cells (5 of 14): the gate fails. Because the per-cell σ_{seed} rests on as few as 4 seeds (3 degrees of freedom on MATH-500), we also report a non-preregistered alternative reading that pools σ_{seed} across the seven validated cells of each benchmark, degrees-of-freedom weighted (pooled values 0.052 and 0.054 for the AIME years, 0.019 for GPQA-Diamond, 0.005 for MATH-500). Under the pooled denominator the hungry-stratum pass rate rises to 7 of 14 (0.500) and the all-cell rate to 10 of 28 (0.357, from 8 of 28). The gate fails its majority bar under either denominator; the frozen per-cell reading remains primary. Screen conditional purity equals the same 35.7% against a 90% bar: fail. On the absolute practitioner bars, the protocol forecast hits 10 of 28 cells and the diagnostic realized-axis readout 13 of 28; every 16k and 32k AIME cell misses low by 9–24 points.

Probe cost fails its preregistered bar, which was denominated in thinking tokens at 20% of the full evaluation: the probe costs 35–38% of the three-cell high suite in thinking tokens on the Qwen3 lineages, and 39.0% on the R1-Distill lineage now that its high suite is complete. Under the alternative GPU-hours accounting the cost is marginal at 21% (6.76 versus 31.5 A100-hours on

8B, similar on 14B), and about 45% of a single 32k cell; GPU-hours beat tokens because attention makes high-budget tokens quadratically expensive [39].

The two criteria that pass require honest annotation. On criterion 6, the selected curve beats the carry-forward baseline, the OLS log-linear baseline (candidate family (a), whose frozen predictions under-predict on all 14 budget-hungry cells), and both preregistered alternatives on the headline stratum: mean noise-normalized errors order as 2.67 selected $<$ 2.95 anchored $<$ 3.04 per-item $<$ 3.58 OLS log-linear $<$ 5.65 carry-forward. The criterion passes against every baseline computable at freeze time, but the mandated 2×-probe baseline was not among them (next paragraph). On criterion 7, the bar is stated on the preregistered noise-normalized axis, where flagged (budget-indifferent) cells show mean error 7.48σ against 2.67σ on screen-passing cells, a 2.8× separation that meets the 2× bar. On raw absolute error the direction *reverses*: flagged cells miss by 0.043 on average against 0.130 for passing cells (ratio 0.33). Both statements are true and the reader should hold both: flat cells have small absolute misses but very small σ_{seed} (MATH-500 pools to 0.005), so two-to-three-point misses become large sigma multiples. The 2.8× figure certifies the screen in noise units on the axis the preregistration specified; it does not validate the decision rule, which is evaluated on its own exposure sets in Section 5.7.

The post-hoc 2×-probe baseline. The plan's third baseline, carrying the measured accuracy at 8192 forward to 16384 and 32768, was never computed before freezing. It is computable today from the landed 8k cells at zero new generation cost, and we report it as a post-hoc, non-preregistered comparison. On the 16 cells where both the 8k anchor and the target had landed at the analysis freeze (Qwen3 models only; the R1-Distill 16k and 32k cells landed after the freeze and are not folded into this post-hoc comparison), the 2×-probe baseline achieves mean absolute error 0.092 against 0.108 for the frozen selected family on the same cells, winning 11 of 16: it is far better on the flat stratum (0.011 versus 0.051 over 8 cells) and slightly worse on the budget-hungry stratum (0.173 versus 0.166 over 8 cells). Criterion 6 therefore stands as preregistered, but the honest overall verdict is mixed: a practitioner willing to pay for one 8k evaluation, which costs roughly the probe again at a higher per-token price, obtains a better point predictor on saturating cells than any curve fit from the ≤4k probe. The frozen curve retains two roles the 2×-probe cannot fill: it is the only forecast available before any high-budget cell exists, and on budget-hungry cells both predictors still under-predict the 32k outcome.

5.3 Calibration

Frozen 90% intervals cover 15 of 28 outcomes (53.6%; 64.3% on the budget-hungry stratum, 42.9% on the flat stratum), well below the 80–97% target. The cause is identifiable: the bootstrap propagates per-cell binomial noise through the *chosen* family, so intervals are appropriately wide where data are noisy (AIME half-widths roughly ±0.12–0.15) but cannot widen for functional-form error, which is the dominant error source. MATH-500 intervals are precise (roughly ±0.03–0.04) and wrong-axis. A calibrated successor needs a form-uncertainty term (for example multi-family mixture intervals) or the decision-layer reframing of Section 5.7. Two construction details bear noting. First, both the measurement CIs and the frozen PIs resample items only, with each item's accuracy averaged over its seeds; any seed-common variance component is treated as zero, an anti-conservative simplification that is second-order here (on MATH-500, $\sigma_{\text{seed}}/\sqrt{4} \leq 0.0041$ against item-bootstrap half-widths of roughly 0.03, and on AIME the 30-item component dominates). Second, no post-hoc widening was applied to the frozen artifacts.

5.4 Error mechanisms

Three mechanisms, each diagnosed from the data, account for the misses.

Curve-shape blindness (budget-hungry cells). The probe cells span the pre-acceleration region, so the LOO-selected sigmoid places its ceiling near the last observed level while the true curve keeps climbing until natural thinking saturates near 12–17k realized tokens. Measured exceeds the protocol forecast Pred@B in all six independent (model, benchmark) budget-hungry curve groups; because the three budgets within a group share one frozen curve, one 30-item set, and overlapping seeds, the honest significance statement is at the group level, a two-sided sign test with $p = 2 \times (1/2)^6 = 0.031$. Descriptively the count is 14 of 14 cells, with under-prediction from +0.03 to +0.24 at nominal x (+0.09 to +0.24 on the 12 Qwen3 AIME cells); the single largest miss is Qwen3-14B AIME-2025 at 32k (measured 0.723 versus predicted 0.485). (On the completed R1-Distill holdout, outside this suite, one 32k buy cell broke the pattern, measuring 0.060 below its frozen forecast; Section 5.7.) The diagnostic readout Pred@$\bar{x}$ makes these cells *worse* (the curve has flattened while accuracy still rises), which is what establishes the miss as genuinely about shape rather than axis.

Axis mismatch (budget-indifferent cells). Qwen3 GPQA and MATH cells stop consuming budget near 4–6.6k realized tokens under a 32k cap, and R1-Distill MATH-500 stalls near 2.7k under its 8k cap. Nominal-x extrapolation then over-predicts by up to 0.10, because it prices in thinking the model never does. The post-hoc diagnostic readout Pred@$\bar{x}$, which requires the held-out cell's measured $\bar{x}$ and is therefore a diagnosis rather than a probe-time forecast, is covered by the measurement CI in 12 of 12 Qwen3 budget-indifferent cells, and all 4 R1-Distill 8k cells are likewise covered at realized x. Here the curve was right and the axis was wrong; the coverage statement uses the measurement CI, a more lenient bar than the preregistered $2\sigma_{\text{seed}}$ gate, and is offered as mechanism attribution, not as a pass.

Overthinking is mild and saturating. MATH-500 accuracy declines by at most 0.023 from its 4k peak and partially recovers by 32k; no cell shows compounding degradation. The inverted-U pathologies of the endogenous-length literature [13] do not materialize under the exogenous knob.

What did not happen also matters: no truncation paradox at the smoke gate, no Control failure, no non-monotone oscillation, and no preregistration integrity failure (Section 3.4).

5.5 Moderators

Error is moderated in the hypothesized directions, and the resulting moderator map is the informative product of the failed gate. *Budget gap:* on the headline stratum, mean absolute error at nominal x grows from 0.083 at the 8k target to 0.177 at 16k, then eases to 0.154 at 32k as the natural-saturation ceiling arrives (thinking plateaus near 12–17k tokens, so the 32k cell is closer to the curve's flat region than the 16k cell); see Figure 2. *Benchmark class* is the dominant moderator, and it is model-dependent rather than benchmark-intrinsic: R1-Distill GPQA is still 30% truncation-bound at 8k where Qwen3 GPQA has stalled near 5k realized tokens (8B stalls at 6.6k, 14B at 5.1k). Class membership must therefore be measured by the probe, not assumed per benchmark. *Model lineage:* the direction and mechanism of every miss replicate across all three lineages; magnitudes differ (R1-Distill thinks more per budget, so its 8k under-predictions are smaller). *Noise floors:* σ_{seed} spans 0.027–0.068 on AIME, 0.015–0.024 on GPQA, and 0.002–0.008 on MATH-500, so the noise-normalized gate is hardest exactly where measurement is most precise, which is why flat MATH cells fail the gate on two-to-three-point misses at realized x.

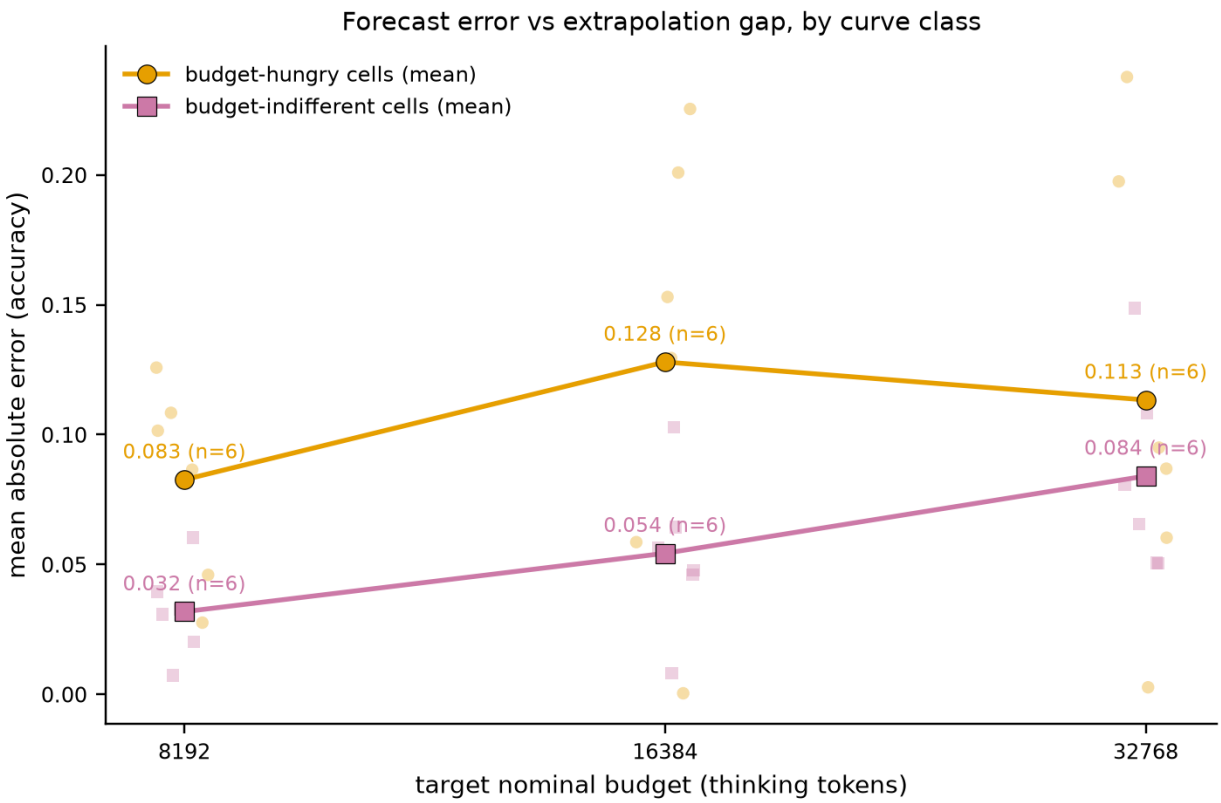

Figure 2: Prediction error versus budget gap, by curve class. Extrapolation error grows with the budget gap on budget-hungry cells and eases slightly at 32k as natural saturation arrives (values in Section 5.5), while budget-indifferent cells show a different failure (axis mismatch) that the post-hoc realized-x diagnostic removes.

5.6 Approach comparison: the aggregate curve beats the per-item model

On the 24 Qwen3 cells, the selected aggregate curve (A) achieves 16/24 CI coverage at realized x against 10/24 for the per-item logistic (B); mean absolute errors are 0.106 versus 0.110 (realized x) and 0.094 versus 0.107 (nominal x). The distributional recipe that wins on the parallel-sampling axis [8] does not transfer: on MATH-500 the monotone per-item ogives compound the same blindness to overthinking saturation and cover 0 of 6 cells against 6 of 6 for A, and on AIME both approaches share the curve-shape failure. The anchored blend (C) never beats A on the headline stratum. The single exception corroborates the preregistration: B wins exactly one cell, R1-Distill MATH-500, the very cell the frozen R1 artifact flagged in advance (its log-linear 0.847 at 8k missed by -0.06 while the frozen per-item 0.759 landed within 0.03 of the measured 0.786). The completed holdout resolved the flagged observable decisively: at 32k the log-linear extrapolation the freeze had called implausible (0.9547) missed the measured 0.8060 by 0.149, while the frozen per-item alternative (0.7912) missed by only 0.015. Across all 28 cells, aggregate MAE is 0.086 at nominal x and 0.097 at realized x, with CI coverage improving from 0.536 to 0.714 when read at realized x: the realized-x readout is not better on raw error (AIME dominates and worsens there) but is better calibrated, and, as a post-hoc diagnostic, it is what isolates the mechanisms of Section 5.4.

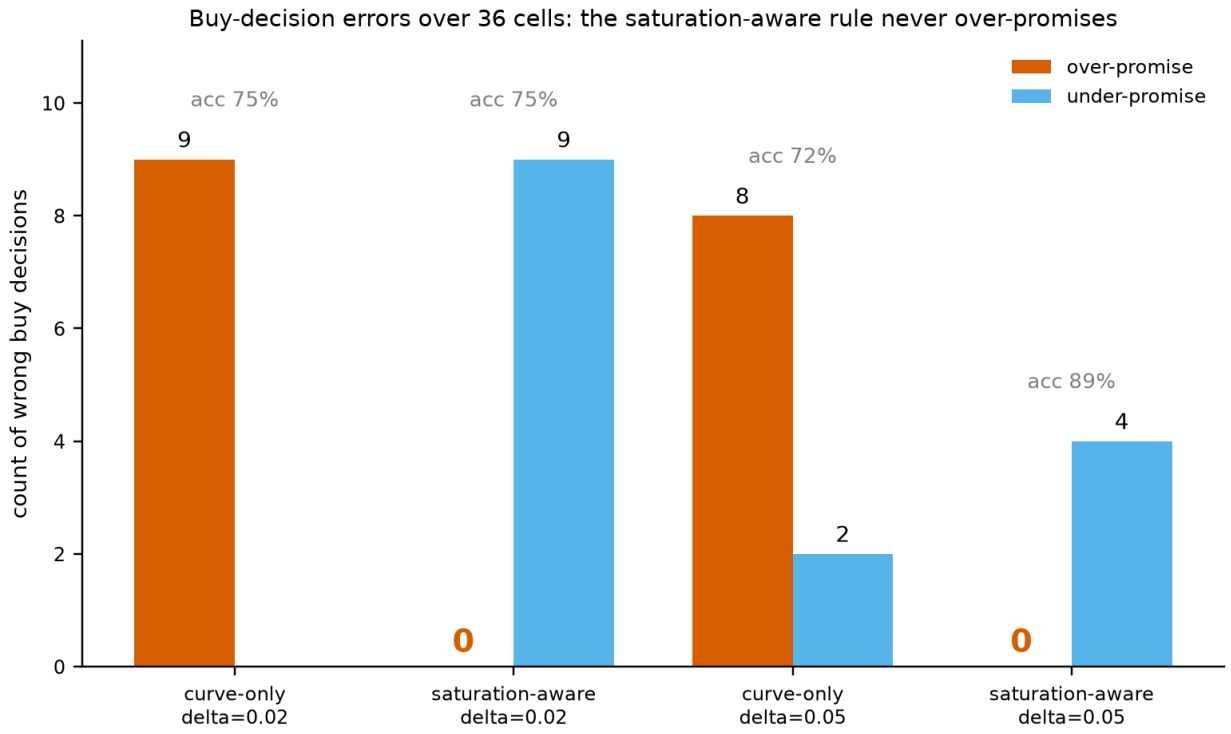

Figure 3: Outcomes of the naive curve-only rule versus the saturation-aware rule at both gain thresholds. Adding the probe-computable saturation check converts every costly over-promise into either a correct decision or a cheap missed upside, at both $\delta = 0.02$ and $\delta = 0.05$, on the validated suite.

5.7 From forecast to purchasing signal

The failed forecast still contains the answer to the practitioner's actual question, provided the exposure is stated honestly: only buy decisions can over-promise, so the relevant denominator is the number of buys, not 28. The naive rule (buy whenever the frozen curve forecasts a gain of at least δ) buys 28 times at $\delta = 0.02$ and 18 times at $\delta = 0.05$ and over-promises 7 and 5 times, with the failures concentrated on budget-indifferent cells where the curve extrapolates a gain the model never collects. The saturation-aware rule of Section 3.5 buys 14 times at $\delta = 0.02$ and 12 times at $\delta = 0.05$ and over-promises zero times in both cases (Table 3, Figure 3).

Zero failures in so few clustered decisions bounds the true rate only loosely: by the rule of three, the 95% upper bound on the over-promise rate is $3/14 = 0.214$ at $\delta = 0.02$ and $3/12 = 0.250$ at $\delta = 0.05$ (pooling the completed holdout into the campaign record gives 18 and 16 buys, all correct, with pooled bounds $3/18 = 16.7\%$ and $3/16 = 18.8\%$; the 6 holdout buys alone bound the rate only at $3/6 = 50\%$), and the buys derive from the six clustered AIME curve groups of Section 5.4, so the effective number of independent decisions is smaller still. Decision accuracy is 0.750 (21 of 28; Wilson 95% CI [0.566, 0.873]) at $\delta = 0.02$ and 0.893 (25 of 28; [0.728, 0.963]) at $\delta = 0.05$.

Every error of the saturation-aware rule is a missed upside: at $\delta = 0.02$ the seven under-promises are all flat-stratum cells with measured gains up to +0.051 (Qwen3-8B GPQA at 16k), while at $\delta = 0.05$ the rule declines two budget-hungry AIME-2025 cells whose forecast gains (0.042, 0.045) fell just below threshold and whose measured gains reached +0.150 and +0.171; the cost asymmetry

is preserved, but the missed upside is not always small.

The screen and rule were constructed on the 24 Qwen3 cells, in full view of the naive rule's failures there, so the 24-cell record is construction, not validation. The rule was fixed roughly three hours before the first R1-Distill high cell landed, so the twelve R1-Distill cells are a genuine holdout, and that holdout is now complete: six buy decisions (the two AIME benchmarks at 8k, 16k, and 32k), all correct at both thresholds, and zero over-promises anywhere in the campaign. At 8k both buy decisions were correct, and the two no-buy decisions were correct at $\delta = 0.05$ but under-promises at $\delta = 0.02$ (GPQA-Diamond +0.023, MATH-500 +0.032 over their probe tops). At 16k the AIME-2024 measurement (0.6292) fell within 0.0005 of its frozen sigmoid forecast (0.6296), AIME-2025 exceeded its forecast gain (+0.196 measured versus +0.137 forecast), and the no-buy cells repeated the 8k pattern (correct at $\delta = 0.05$, under-promises at $\delta = 0.02$: +0.017 GPQA-Diamond, +0.043 MATH-500). The final 32k cell closed the campaign with both buy decisions correct by wide margins (AIME-2024 gained +0.271 against a forecast gain of +0.331; AIME-2025 gained +0.206 against +0.209) and carries the record's one substantive forecast shortfall: the AIME-2024 measurement (0.6750) fell 0.060 below its frozen forecast (0.7353), because R1-Distill's natural thinking saturates near 10.7k realized tokens, well short of what the frozen sigmoid assumed. The 32k no-buy decisions were correct on GPQA-Diamond at both thresholds (measured gain +0.010) and under-promises on MATH-500 at both (measured gain +0.052, the largest missed upside of the campaign at $\delta = 0.02$). Six holdout buys still cannot tightly bound an error rate. The one-sidedness result strengthens the signal beyond the binary decision, with its scope now stated exactly: on the preregistered suite the frozen forecast under-predicted in all six curve groups (14 of 14 budget-hungry cells), and on the completed holdout the measured score beat the frozen forecast or matched it to within 0.003 on five of six buy cells and fell 0.060 short on the sixth. When the rule said buy, the frozen forecast functioned as an approximate lower bound on what the practitioner actually got, not a guaranteed one; on the one exception the buy was still correct by a wide margin, which is the record's point: the decision is robust even where the point forecast is not.

Table 1: The 28-cell validation suite. Measured accuracy with 95% CI (item-level percentile bootstrap, 1000 replicates, each item's accuracy pooled over its seeds), mean realized thinking tokens $\bar{x}$, the frozen protocol forecast Pred@B (curve read at the nominal budget; the preregistered object, available at probe time), and the post-hoc diagnostic readout Pred@$\bar{x}$ (curve read at the target cell's *measured* $\bar{x}$; not available at probe time). The 90% PI column is the frozen 90% prediction interval for Pred@B. Asterisks mark measured values outside the frozen 90% PI (13 of 28; coverage 53.6%). In this suite every AIME cell is budget-hungry (screen-passing) and every GPQA-Diamond and MATH-500 cell is budget-indifferent (screen-flagged). The starred Qwen3-14B MATH-500 cell at 8192 is a boundary case: at full precision the measured value falls 0.0003 below the PI lower bound (0.8235 versus 0.8238), a miss invisible at the displayed rounding. Because the frozen prediction intervals were built from 200 bootstrap replicates, the Monte Carlo standard error of the PI endpoint exceeds this 0.0003 margin, so that cell's miss-or-cover status is within bootstrap resolution and empirical coverage is 15-or-16 of 28; the calibration verdict (fail against the 80–97% target) is unchanged at either count. R1-Distill cells at 16k and 32k are excluded from the preregistered suite by construction (the suite was fixed at the analysis freeze, before those cells landed) and are reported in the holdout paragraphs of Section 5.7.

Benchmark	B	Measured	95% CI	$\bar{x}$	Pred@B	Pred@$\bar{x}$	90% PI
Qwen3-8B							
AIME-2024	8192	0.575	[0.429, 0.715]	7433	0.489	0.475	[0.376, 0.620]
AIME-2024	16384	0.715	[0.573, 0.831]	11118	0.585	0.531	[0.445, 0.723]
AIME-2024	32768	0.762	[0.631, 0.881]	13708	0.676	0.561	[0.519, 0.813]
AIME-2025	8192	0.473	[0.321, 0.646]	7544	0.365	0.358	[0.234, 0.480]
AIME-2025	16384	0.621*	[0.473, 0.763]	12479	0.420	0.398	[0.286, 0.558]
AIME-2025	32768	0.675*	[0.523, 0.808]	16650	0.478	0.421	[0.342, 0.627]
GPQA-D	8192	0.572	[0.517, 0.628]	5470	0.565	0.547	[0.509, 0.614]
GPQA-D	16384	0.587	[0.527, 0.642]	6379	0.595	0.554	[0.533, 0.649]
GPQA-D	32768	0.574	[0.516, 0.626]	6587	0.625	0.555	[0.557, 0.686]
MATH-500	8192	0.820*	[0.789, 0.849]	3885	0.852	0.836	[0.821, 0.880]
MATH-500	16384	0.811*	[0.778, 0.844]	4566	0.867	0.840	[0.830, 0.902]
MATH-500	32768	0.817*	[0.782, 0.850]	4814	0.882	0.841	[0.839, 0.924]
Qwen3-14B							
AIME-2024	8192	0.627	[0.490, 0.765]	7198	0.526	0.507	[0.400, 0.654]
AIME-2024	16384	0.777*	[0.646, 0.904]	10629	0.624	0.563	[0.489, 0.756]
AIME-2024	32768	0.808	[0.683, 0.923]	12737	0.713	0.589	[0.575, 0.841]
AIME-2025	8192	0.500	[0.358, 0.665]	7431	0.374	0.367	[0.241, 0.500]
AIME-2025	16384	0.654*	[0.508, 0.783]	12086	0.429	0.405	[0.281, 0.572]
AIME-2025	32768	0.723*	[0.579, 0.863]	15474	0.485	0.424	[0.322, 0.658]
GPQA-D	8192	0.596	[0.537, 0.658]	4556	0.636	0.606	[0.588, 0.693]
GPQA-D	16384	0.623	[0.566, 0.683]	5048	0.670	0.611	[0.617, 0.735]
GPQA-D	32768	0.624*	[0.567, 0.684]	5103	0.705	0.612	[0.641, 0.773]
MATH-500	8192	0.824*	[0.788, 0.854]	3466	0.854	0.839	[0.824, 0.886]
MATH-500	16384	0.821*	[0.787, 0.854]	3986	0.867	0.842	[0.831, 0.907]
MATH-500	32768	0.829*	[0.796, 0.859]	4095	0.879	0.842	[0.839, 0.927]
DeepSeek-R1-Distill-Qwen-14B							
AIME-2024	8192	0.538	[0.406, 0.673]	6464	0.510	0.468	[0.411, 0.637]
AIME-2025	8192	0.404	[0.262, 0.546]	6621	0.358	0.338	[0.223, 0.467]
GPQA-D	8192	0.589	[0.534, 0.645]	5265	0.609	0.585	[0.556, 0.659]
MATH-500	8192	0.787*	[0.751, 0.818]	2737	0.847	0.761	[0.812, 0.879]

Table 2: Preregistered success criteria and verdicts. The point-prediction and calibration gates fail; the baseline-margin and screen-validation criteria pass.

Criterion	Preregistered bar	Result	Verdict
1. Noise-normalized gate	within $2\sigma_{\text{seed}}$ for majority of well-behaved cells	best approach passes 35.7% (5/14) of budget-hungry cells (mean 2.67σ); 50.0% (7/14) under the post-hoc pooled-σ reading	fail
2. Screen conditional purity	$\geq 90\%$ of screen-passing cells meet gate	35.7% (same cell set as criterion 1)	fail
3. No trivial-cell padding	stratified reporting; flat cells excluded from headline	done (14 hungry / 14 flat)	met
4. Calibration	frozen 90% PIs cover 80–97%	53.6% (hungry 64.3%, flat 42.9%)	fail
5. Probe cost	$\leq$ 20% of full-evaluation thinking tokens	35–39% tokens; 21% GPU-hours and ∼45% vs. one 32k cell as alternative accountings	fail
6. Baseline margin	best method beats every naive baseline (mean nn-error)	headline: $2.67 < 2.95 < 3.04 < 3.58 < 5.65$ vs. anchored, per-item, OLS log-linear, and carry-forward; post-hoc 2×-probe baseline beats the frozen curve on 11/16 comparable cells (Section 5.2); on flat stratum carry-forward wins (2.63 vs. 3.01)	pass vs. frozen-time baselines; mixed vs. post-hoc 2×-probe
7. Screen validation	flagged cells $\geq 2\times$ error of passing cells (noise-normalized)	2.8× noise-normalized; raw-error ratio 0.33 (direction reverses)	pass on preregistered axis

Table 3: Decision-rule comparison over all 28 validated cells: should the practitioner pay for budget $B > 4096$ to gain at least δ? Buy decisions are the exposure set for over-promises (buy, but the measured gain falls short: the costly error); no-buy decisions are the exposure set for under-promises. The saturation-aware rule makes zero over-promises at both thresholds; with zero failures the 95% upper bound (UB) on the true over-promise rate is the rule-of-three bound $3/n_{\text{buy}}$. Accuracy intervals are Wilson 95%. Of the saturation-aware buys, 2 at each δ are R1-Distill holdout decisions.

Rule	δ	Buys	No-buys	Over	Under	Accuracy [95% CI]	Over-promise rate
Naive (curve only)	0.02	28	0	7	0	0.750 [0.566, 0.873]	7/28
Saturation-aware	0.02	14	14	0	7	0.750 [0.566, 0.873]	0/14 (95% UB 0.214)
Naive (curve only)	0.05	18	10	5	2	0.750 [0.566, 0.873]	5/18
Saturation-aware	0.05	12	16	0	3	0.893 [0.728, 0.963]	0/12 (95% UB 0.250)

6 Discussion

Why lead with a failed gate. The preregistration is what makes both halves of the finding credible. Had the fits been free to adjust after the high-budget cells landed, the 35.7% gate pass would likely have become a substantially more favorable number, and the one-sidedness of the misses, the fact that carries all the practical value, could have been fit away. The frozen artifacts instead expose a stable structure: low-budget probes cannot see the shape of a curve above the region they sample (an instance of the general result that breaks above the fit range are unextrapolatable [31]), but they can see, in their own realized-token statistics, whether the model has anything left to gain.

What a practitioner should do. The protocol our data support is: run the probe grid (budgets 512–4096 plus the $B = 0$ anchor, with seed replication and the Control check); fit and freeze the curve families on realized tokens; compute the saturation check. If the probe is not saturating its cap, do not buy more budget, and treat the probe-top accuracy as the working estimate: on our budget-indifferent cells the measured high-budget score never moved more than +0.052 from the probe top. The realized-x curve readout of Section 5.4 is not part of this recipe, because it requires the target cell's measured realized tokens; it is a post-hoc diagnostic, not probe-time guidance. If the probe is saturating and the frozen curve forecasts a gain of at least the practitioner's threshold, buy: across the campaign this signal never over-sold in any of its buy decisions (14 in the preregistered suite, 6 on the completed holdout), and the point forecast functioned as an approximate floor, not an estimate, of the high-budget score, with one holdout exception where the measured score fell 0.060 below the frozen forecast while the buy remained correct by a wide margin. What the probe cannot do is replace the high-budget run when an accurate point value is itself the deliverable.

Benchmark class is a measurement, not a property, and the screen can misread it. The same benchmark can be budget-hungry for one lineage and budget-indifferent for another at the same nominal budget (GPQA-Diamond; numbers in Section 5.5), so any protocol that hard-codes benchmark classes can misclassify. The probe's realized-token statistics recover class membership cheaply, but not infallibly, and R1-Distill GPQA is the instructive failure: the probe realized 82% of the 4k cap (3370 of 4096 tokens), below the 90% threshold, so the screen classified the cell budget-indifferent, yet the high-budget cell was still 30% truncation-bound at 8k and gained +0.023, making the classification an under-promise error at $\delta = 0.02$. The regime difference was revealed by high-budget truncation, which the probe-only screen does not observe. In our suite this screen failure mode was confined to the cheap direction (a missed upside, not an over-sold gain), consistent with the screen's conservative bias, but that confinement is an empirical observation, not a guarantee.

Is the one-sided bound a truncation artifact? A natural objection is that fitting on heavily truncated probe cells would mechanically produce under-prediction: the 4k AIME probe cells truncate 78–100% of responses, so the probe accuracies may be depressed relative to what untruncated reasoning would score at the same realized length, and a curve fit to depressed points would freeze its ceiling too low. What our data do establish is that the held-out measurements themselves are not truncation-suppressed: at $B = 32768$ the four Qwen3 AIME cells truncate only 2–10% of responses (0.021–0.098), so the measured 32k accuracies approximate the native ceiling. What the data cannot establish is why the frozen curve sits below them: probe-side truncation depression of

the fit cells and intrinsic curve-shape blindness both predict the same persistent under-prediction, largest exactly where truncation has relaxed, and under the cap knob the two are observationally equivalent, because truncation relief and gains from more thinking are the same event; only the unrun native-versus-cap ablation could separate them (Section 7). For the practitioner the operational consequence is identical under either mechanism: in the preregistered suite the frozen forecast under-predicted on every budget-hungry cell, and on the completed holdout the measured score beat the frozen forecast or matched it to within 0.003 on five of six buy cells while the forecast overshot once by 0.060, so the forecast functioned as an approximate floor, not a guaranteed one, whichever cause depressed its ceiling.

Where the per-item recipe stalls. The parallel-axis prediction literature [8] suggests per-item distributional modeling should dominate aggregate curves. On the sequential axis it does not: per-item monotone ogives inherit, and on saturating benchmarks compound, the same functional-form blindness. The one preregistered exception (R1-Distill MATH-500, where the aggregate family was flagged implausible before freezing) suggests per-item modeling is worthwhile precisely when the aggregate family is known to be misspecified, a diagnosable condition rather than a default.

7 Limitations

Internal validity. AIME cells rest on 30 items; even with 16 seeds, σ_{seed} reaches 0.068 and 95% CIs span roughly ± 0.14, so several individual AIME cells technically cover their predictions even where the systematic one-sided pattern (all six curve groups, cluster-level $p = 0.031$) establishes the miss. The stringency of the noise-normalized gate itself rests on per-cell σ_{seed} estimates with as few as 3 degrees of freedom on MATH-500, so part of the gate's cross-benchmark stringency variation is sampling noise in the denominator rather than real noise-floor differences; the pooled-σ reading of Section 5.2 bounds the effect (the gate fails either way). The frozen 90% intervals are miscalibrated for a diagnosed reason (no functional-form uncertainty term, Section 5.3); we report the frozen intervals rather than post-hoc widened ones. The eight R1-Distill cells at 16k and 32k were preregistered but landed after the analysis freeze; the 28-cell suite stands without them, and they are reported as the completed holdout. Relatedly, the saturation screen and decision rule were developed on the 24 Qwen3 cells and are held out only on the 12 R1-Distill cells, which landed with all 6 holdout buy decisions correct, so the zero-over-promise record (0 of 14 and 0 of 12 buys on the suite; 0 of 18 and 0 of 16 pooling in the holdout) remains largely in-sample, and its rule-of-three bounds (pooled $3/18 = 16.7\%$ at $\delta = 0.02$, holdout-only $3/6 = 50\%$; Section 5.7) should be read with that caveat. Contamination of public competition benchmarks cannot be ruled out post hoc [40]; we mitigate by reporting AIME 2024 and 2025 as separate strata. The two years show the same miss direction and mechanisms, but AIME-2024 sits +0.085 to +0.133 above AIME-2025 at every paired high-budget cell, a stable level gap consistent both with contamination of the older set and with a harder 2025 problem set; our design cannot adjudicate between these explanations, so no cross-year level claim is made. MATH-500 is plausibly present in supervised fine-tuning corpora, which is one reason its regime is saturated by design.

External validity. All results are at model scales of 14B parameters and below; a planned Qwen3-32B extension was cut (it does not fit a single A100 80GB in bf16 at 40k context, and the campaign was restricted to single-GPU serving), so no claim is made above 14B. The suggestive trend that

larger models saturate earlier on GPQA (Section 5.5) rests on two points and is flagged as a question, not a finding. Each lineage was tested with a single knob implementation (the shared cap harness of Section 3.1, with lineage-specific $B = 0$ anchors); the planned native-versus-truncation ablation on a single model was not run, so differences between cap-budget and native-budget curves are unquantified, knob-mechanism effects and lineage effects are partially confounded, and a practitioner with a natively trained budget control should treat our curves as the cap-construct response only, though the cross-lineage replication of every qualitative result bounds the concern within the cap construct. In addition, all budget-hungry cells in the validated suite are AIME cells: no budget-hungry non-AIME cell exists in our data, so the buy-arm record rests on one benchmark family with 60 unique items. A planned relative-budget normalization ablation (curve collapse at x expressed as a fraction of natural thinking length) was also not run; both ablations are recorded as future work.

Construct validity. Accuracy at nominal budget B is the practitioner-facing construct, but the more physically grounded quantity is accuracy at realized thinking tokens; the divergence between the two is itself one of our findings, and any deployment of the protocol must decide which axis its consumer cares about. The 90%-of-cap screen threshold and the two gain thresholds (0.02, 0.05) were fixed without a sensitivity sweep, so the robustness of the 14/14 split and of the decision record to these choices is unquantified; the observed saturation statistic happens to leave an empty margin around 0.9 (flat cells at or below 86.4% of cap, hungry cells at or above 92.8%), so any threshold in that margin reproduces the same split in-sample, but a cell landing inside the margin is unadjudicated by our data. Rank preservation across models at high budget was planned as a metric but not computed. All serving ran one stack (vLLM 0.10.1, temperature 0.6, top-p 0.95) on A100 hardware; curve levels, though plausibly not the saturation structure, may shift under other serving configurations. Finally, the probe is cheap only under the GPU-hours accounting (full figures in Section 5.2); with the R1-Distill high suite complete, that lineage's probe token fraction is 0.390, in line with the Qwen3 lineages and likewise above the preregistered 20% bar.

8 Availability

The artifacts of this study, including the evaluation harness code, per-response generation records with realized thinking-token counts, the frozen and hashed preregistration files, and the analysis scripts, are available from the authors on request.

References

[1] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314 [cs.LG], 2024. URL https://arxiv.org/abs/2408.03314.

[2] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang,

Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL], 2025. URL https://arxiv.org/abs/2505.09388.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL], 2025. URL https://arxiv.org/abs/2501.12948.

[4] Pranjal Aggarwal and Sean Welleck. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697 [cs.CL], 2025. URL https://arxiv.org/abs/2503.04697.

[5] Baohao Liao, Hanze Dong, Yuhui Xu, Doyen Sahoo, Christof Monz, Junnan Li, and Caiming Xiong. Fractured Chain-of-Thought Reasoning. arXiv:2505.12992 [cs.LG], 2025. URL https://arxiv.org/abs/2505.12992.

[6] Ziqian Bi, Lu Chen, Junhao Song, Hongying Luo, Enze Ge, Junmin Huang, Tianyang Wang, Keyu Chen, Chia Xin Liang, Zihan Wei, Huafeng Liu, Chunjie Tian, Jibin Guan, Joe Yeong, Yongzhi Xu, Peng Wang, Xinyuan Song, and Junfeng Hao. Exploring Efficiency Frontiers

of Thinking Budget in Medical Reasoning: Scaling Laws between Computational Resources and Reasoning Quality. arXiv:2508.12140 [cs.CL], 2025. URL https://arxiv.org/abs/2508.12140.

[7] Austin R. Ellis-Mohr, Anuj K. Nayak, and Lav R. Varshney. A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search. arXiv:2507.00004 [cs.LG], 2025. URL https://arxiv.org/abs/2507.00004.

[8] Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient Prediction of Pass@k Scaling in Large Language Models. arXiv:2510.05197 [cs.AI], 2025. URL https://arxiv.org/abs/2510.05197.

[9] Sijie Li, Shanda Li, Haowei Lin, Weiwei Sun, Ameet Talwalkar, and Yiming Yang. Spend Less, Fit Better: Budget-Efficient Scaling Law Fitting via Active Experiment Selection. arXiv:2604.22753 [cs.LG], 2026. URL https://arxiv.org/abs/2604.22753.

[10] Hanlin Zhang, Jikai Jin, Vasilis Syrgkanis, and Sham Kakade. Prescriptive Scaling Reveals the Evolution of Language Model Capabilities. arXiv:2602.15327 [cs.LG], 2026. URL https://arxiv.org/abs/2602.15327.

[11] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance. arXiv:2405.10938 [cs.LG], 2024. URL https://arxiv.org/abs/2405.10938.

[12] Mohammad Ali Alomrani, Yingxue Zhang, Derek Li, Qianyi Sun, Soumyasundar Pal, Zhanguang Zhang, Yaochen Hu, Rohan Deepak Ajwani, Antonios Valkanas, Raika Karimi, Peng Cheng, Yunzhou Wang, Pengyi Liao, Hanrui Huang, Bin Wang, Jianye Hao, and Mark Coates. Reasoning on a Budget: A Survey of Adaptive and Controllable Test-Time Compute in LLMs. arXiv:2507.02076 [cs.AI], 2025. URL https://arxiv.org/abs/2507.02076.

[13] Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When More is Less: Understanding Chain-of-Thought Length in LLMs. arXiv:2502.07266 [cs.AI], 2025. URL https://arxiv.org/abs/2502.07266.

[14] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? arXiv:2502.12215 [cs.LG], 2025. URL https://arxiv.org/abs/2502.12215.

[15] Daisuke Nohara, Taishi Nakamura, and Rio Yokota. On the Optimal Reasoning Length for RL-Trained Language Models. arXiv:2602.09591 [cs.CL], 2026. URL https://arxiv.org/abs/2602.09591.

[16] Marthe Ballon, Andres Algaba, and Vincent Ginis. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. arXiv:2502.15631 [cs.LG], 2025. URL https://arxiv.org/abs/2502.15631.

[17] Mathematical Association of America. American Invitational Mathematics Examination (AIME) 2024. Competition problem set; evaluation snapshot: HuggingFace dataset `HuggingFaceH4/aime_2024`, 2024. URL https://huggingface.co/datasets/HuggingFaceH4/aime_2024.

[18] Mathematical Association of America. American Invitational Mathematics Examination (AIME) 2025. Competition problem set; evaluation snapshot: HuggingFace dataset `math-ai/aime25`, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

[19] 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. arXiv:2103.03874 [cs.LG], 2021. URL https://arxiv.org/abs/2103.03874.

[20] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. arXiv:2305.20050 [cs.LG], 2023. URL https://arxiv.org/abs/2305.20050.

[21] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022 [cs.AI], 2023. URL https://arxiv.org/abs/2311.12022.

[22] Vidhisha Balachandran, Jingya Chen, Lingjiao Chen, Shivam Garg, Neel Joshi, Yash Lara, John Langford, Besmira Nushi, Vibhav Vineet, Yue Wu, and Safoora Yousefi. Inference-Time Scaling for Complex Tasks: Where We Stand and What Lies Ahead. arXiv:2504.00294 [cs.LG], 2025. URL https://arxiv.org/abs/2504.00294.

[23] 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. arXiv:2501.19393 [cs.CL], 2025. URL https://arxiv.org/abs/2501.19393.

[24] Michael Kleinman, Matthew Trager, Alessandro Achille, Wei Xia, and Stefano Soatto. e1: Learning Adaptive Control of Reasoning Effort. arXiv:2510.27042 [cs.AI], 2025. URL https://arxiv.org/abs/2510.27042.

[25] Bo Zhao, Berkcan Kapusuzoglu, Kartik Balasubramaniam, Sambit Sahu, Supriyo Chakraborty, and Genta Indra Winata. Optimizing Reasoning Efficiency through Prompt Difficulty Prediction. arXiv:2511.03808 [cs.LG], 2025. URL https://arxiv.org/abs/2511.03808.

[26] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941 [cs.AI], 2025. URL https://arxiv.org/abs/2506.06941.

[27] James Xu Zhao, Bryan Hooi, and See-Kiong Ng. Test-Time Scaling in Reasoning Models Is Not Effective for Knowledge-Intensive Tasks Yet. arXiv:2509.06861 [cs.AI], 2025. URL https://arxiv.org/abs/2509.06861.

[28] Wang Yang, Debargha Ganguly, Xinpeng Li, Chaoda Song, Shouren Wang, Vikash Singh, Vipin Chaudhary, and Xiaotian Han. Mid-Think: Training-Free Intermediate-Budget Reasoning via Token-Level Triggers. arXiv:2601.07036 [cs.CL], 2026. URL https://arxiv.org/abs/2601.07036.

[29] Guojun Wu. It's Not That Simple. An Analysis of Simple Test-Time Scaling. arXiv:2507.14419 [cs.LG], 2025. URL https://arxiv.org/abs/2507.14419.

[30] Ian Su, Gaurav Purushothaman, Jey Narayan, Ruhika Goel, Kevin Zhu, Sunishchal Dev, Yash More, and Maheep Chaudhary. Broken Chains: The Cost of Incomplete Reasoning in LLMs. arXiv:2602.14444 [cs.LG], 2026. URL https://arxiv.org/abs/2602.14444.

[31] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken Neural Scaling Laws. arXiv:2210.14891 [cs.LG], 2022. URL https://arxiv.org/abs/2210.14891.

[32] Dongwoo Lee, Dong Bok Lee, Steven Adriaensen, Juho Lee, Sung Ju Hwang, Frank Hutter, Seon Joo Kim, and Hae Beom Lee. Bayesian Neural Scaling Law Extrapolation with Prior-Data Fitted Networks. arXiv:2505.23032 [cs.LG], 2025. URL https://arxiv.org/abs/2505.23032.

[33] Peiyao Cai, Chengyu Cui, Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Kean Ming Tan, and Gongjun Xu. A Latent Variable Framework for Scaling Laws in Large Language Models. arXiv:2512.06553 [stat.AP], 2025. URL https://arxiv.org/abs/2512.06553.

[34] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. arXiv:2402.14992 [cs.CL], 2024. URL https://arxiv.org/abs/2402.14992.

[35] Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How Benchmark Prediction from Fewer Data Misses the Mark. arXiv:2506.07673 [cs.LG], 2025. URL https://arxiv.org/abs/2506.07673.

[36] David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation. arXiv:2508.13144 [cs.CL], 2025. URL https://arxiv.org/abs/2508.13144.

[37] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004 [cs.AI], 2023. URL https://arxiv.org/abs/2304.15004.

[38] Rosie Zhao, Tian Qin, David Alvarez-Melis, Sham Kakade, and Naomi Saphra. Random Scaling of Emergent Capabilities. arXiv:2502.17356 [cs.LG], 2025. URL https://arxiv.org/abs/2502.17356.

[39] Ranajoy Sadhukhan, Zhuoming Chen, Haizhong Zheng, Yang Zhou, Emma Strubell, and Beidi Chen. Kinetics: Rethinking Test-Time Scaling Laws. arXiv:2506.05333 [cs.LG], 2025. URL https://arxiv.org/abs/2506.05333.

[40] Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On The Fragility of Benchmark Contamination Detection in Reasoning Models. arXiv:2510.02386 [cs.CR], 2025. URL https://arxiv.org/abs/2510.02386.

[41] Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. Demystifying Hybrid Thinking: Can LLMs Truly Switch Between Think and No-Think? arXiv:2510.12680 [cs.LG], 2025. URL https://arxiv.org/abs/2510.12680.

[42] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932 [cs.LG], 2024. URL https://arxiv.org/abs/2404.09932.

[43] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180 [cs.LG], 2023. URL https://arxiv.org/abs/2309.06180.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL], 2020. URL https://arxiv.org/abs/1910.03771.

[45] Hynek Kydlíček. Math-Verify: A Robust Mathematical Expression Evaluation System. GitHub repository, version 0.7.0, 2025. URL https://github.com/huggingface/math-verify.