# Training Quality Determines Efficiency Boundaries in Test-Time Reasoning

**MD Azizul Hakim**
Department of Computer Science and Technology
Bangladesh Sweden Polytechnic Institute, Chittagong, Bangladesh
azizulhakim8291@gmail.com

## Abstract

Training quality, rather than parameter count or inference budget, is the strongest determinant of reasoning efficiency observed across the large language models evaluated here. Evaluating 50 language models (0.5B–685B parameters) across six reasoning benchmarks totalling over 67,000 assessments, this study characterises when additional inference tokens are beneficial, redundant or actively harmful. Standard models exhibit near-universal accuracy convergence at approximately 1,000 tokens across architectures and parameter counts, a boundary that persists under temperature sampling. Causal intervention experiments show that constraining generation length improves GSM8K accuracy by 15.4 percentage points, with gains replicating across all eight randomly sampled model–dataset conditions, identifying over-generation rather than capacity exhaustion as the primary efficiency bottleneck. Among 19 reasoning-specialised models, training methodology yields a 5.0–18.8-fold token-efficiency gap at matched parameter scales that widens with generation budget, exceeding the efficiency gains from ten-fold parameter increases; the advantage persists under self-consistency decoding. Graduate-level evaluation reveals that arithmetic benchmark performance does not predict reasoning capability on complex tasks ($r = -0.027$, $P = 0.915$, $n = 18$; power sufficient to detect $r \geq 0.6$ but not to exclude moderate correlations), exposing limitations invisible to standard evaluation practice. Mechanistic profiling reveals two orthogonal dimensions of training quality—quality control and decomposition efficiency—that jointly predict efficiency profiles and correctly assign all ten held-out models from five model families to their efficiency tier, providing a measurable framework linking training decisions to inference-time behaviour.

**Full Diagnostic Code**
**All Reasoning Models Mechanistic Dataset**

## 1 Introduction

Scaling laws establish that large language model performance improves predictably with increasing parameters and training data Kaplan et al. (2020); Hoffmann et al. (2022), and recent systematic analysis demonstrates that training stage decisions—pre-training composition, continued pre-training, and post-training methodology—produce significant capability differences across model scales Qi et al. (2025). A parallel assumption—that allocating additional tokens during inference monotonically improves reasoning—now drives the test-time compute paradigm, with models such as o1 Jaech et al. (2024), DeepSeek-R1 Guo et al. (2025) and Gemini 2.5 Comanici et al. (2025) consuming substantially longer reasoning chains to solve complex tasks. Recent work on pre-training efficiency has formalised the densing law, showing that capability density doubles approximately every 3.5 months Xiao et al. (2025), yet no analogous framework characterises when additional inference tokens

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

are beneficial, redundant or actively harmful. This asymmetry leaves a central question unresolved: *do efficiency boundaries exist for test-time reasoning, and if so, what determines them?*

Reasoning-specialised models routinely generate 10–100× more tokens per query than standard models Snell et al. (2024); Wu et al. (2024); if many of these tokens add no accuracy, current deployment is systematically wasteful, yet existing studies remain fragmented across individual model families. (Extended related work in Appendix A.)

This study addresses the question through systematic evaluation of 50 language models (0.5B–685B parameters) across six reasoning benchmarks, totalling over 67,000 individual assessments Cobbe et al. (2021a); Clark et al. (2019, 2018); Talmor et al. (2019); Hendrycks et al. (2021); Rein et al. (2024). Initial analysis across 31 standard models reveals that longer reasoning chains correlate negatively with accuracy ($\rho = -0.093$, $P < 10^{-25}$; Appendix B and C), motivating controlled investigation of efficiency boundaries. The study combines token ablation experiments (systematically varying maximum generation length), causal intervention experiments (constraining verbosity to test over-generation), and mechanistic profiling of 19 reasoning-specialised models (13 across four datasets, six additional for framework validation) Jaech et al. (2024); Guo et al. (2025); Cai et al. (2024); He et al. (2025). We use *training quality* to denote the composite outcome of training methodology decisions (reward design, data curation, optimisation procedure) as measured through observable output properties.

First, standard models exhibit near-universal accuracy convergence at approximately 1,000 generated tokens, with 87.5% of model–task conditions showing zero improvement beyond this boundary (8 genuine convergence, 12 flat from the lowest limit, 1 declining; Appendix D.1.1). Second, causal intervention experiments indicate that constraining generation length improves GSM8K accuracy by 15.4 percentage points ($P = 0.010$), with gains replicating across all eight randomly sampled model–dataset conditions, identifying over-generation rather than capacity exhaustion as the primary bottleneck. Third, among reasoning-specialised models, training methodology yields a 5.0–18.8-fold token-efficiency gap at matched parameter scales, exceeding the efficiency gains from ten-fold parameter increases. Fourth, mechanistic profiling reveals two orthogonal dimensions of training quality—quality control and decomposition efficiency—that jointly predict efficiency profiles and correctly assign all ten held-out models from five model families to their efficiency tier. The convergence boundary and the over-generation effect both hold under temperature sampling, and the training-quality efficiency advantage persists under self-consistency decoding (Appendix D.11 and D.12). These patterns give rise to four corollaries:

**Corollary 1.** *For standard models, token budgets beyond approximately 1,000 tokens yield zero accuracy improvement across 87.5% of tested conditions, representing budget allocation with no measurable return.*

**Corollary 2.** *The densing law Xiao et al. (2025) characterises training-phase efficiency; the present findings characterise inference-phase efficiency. Together these frameworks span the optimisation landscape from training to deployment.*

**Corollary 3.** *Extended generation degrades performance through unfocused elaboration rather than capacity exhaustion: blinded two-annotator analysis (Cohen's $\kappa = 0.886$) shows 76% of failures under unconstrained budgets involve abandonment of initially correct reasoning in favour of unfocused elaboration (Appendix D.7).*

**Corollary 4.** *Arithmetic benchmark performance does not predict graduate-level reasoning performance ($r = -0.027$, $P = 0.915$, $n = 18$; power sufficient to detect $r \geq 0.6$). Standard benchmarks alone may mask fundamental reasoning limitations.*

## 2 Methods

### 2.1 Evaluation framework

This study comprised three experimental phases totalling over 67,000 evaluations across 50 language models. Phase 1 assessed 31 standard models (0.5B–405B parameters) across 11 model families (Llama, Qwen, Mistral, Gemma, Phi, Yi, GPT, StableLM, DeepSeek, Kimi, Gemini) and five reasoning benchmarks: GSM8K Cobbe et al. (2021a) (grade-school mathematics), BoolQ Clark et al. (2019) (boolean question answering), ARC-Easy Clark et al. (2018) (science reasoning),

CommonsenseQA Talmor et al. (2019) (commonsense reasoning), and MMLU-STEM Hendrycks et al. (2021) (19 STEM subjects), sampling 300 problems per dataset (46,500 evaluations). Phase 2 conducted systematic token ablation on eight representative models spanning a 405-fold parameter range (1B–405B) at four token limits (300, 500, 1,000, 2,000 tokens; $n = 50$ per condition) across three datasets (GSM8K, BoolQ, ARC-Easy), totalling 4,800 evaluations (capped at 2,000 tokens, consistent with standard models' practical generation range). Phase 3 evaluated 13 reasoning-specialised models (1.5B–685B parameters) representing distillation Guo et al. (2025), reinforcement learning He et al. (2025), and Monte Carlo tree search Silver et al. (2016) training methodologies across four datasets including GPQA Rein et al. (2024) (graduate-level science questions; main split, 50 stratified samples) at up to six token limits (300–8,000 tokens; small models at 300–4,000, large models at 300–8,000; $n = 50$ per condition), producing approximately 17,000 evaluations. Six further reasoning models (7B–229B) were evaluated on GSM8K and GPQA under identical protocols for framework validation and correlation analysis (Appendix H). Open-weight models were evaluated with HuggingFace Transformers Wolf et al. (2020) in FP16 precision with greedy decoding (`do_sample = False`); Llama-3.1-405B, Gemini-2.0-Flash, Kimi2.5-1T, GLM-4.7-355B, the MiniMax models and DeepSeek-R1-685B were queried through provider APIs with greedy (temperature-0) decoding and identical zero-shot prompts and token limits. No model-specific prompt engineering was applied (Appendix D.5). Answers were scored with a single hierarchical extraction pipeline applied identically to all models and conditions (Appendix D.10.2). To test sensitivity to decoding strategy, three robustness experiments used temperature sampling ($T = 0.6$, top-$p = 0.95$; $n = 50$): (i) token ablation at 1,000 and 2,000 tokens for three Phase 2 models on the same datasets and problems; (ii) the causal intervention for Qwen2.5-32B and Llama-3.3-70B on GSM8K and MMLU-STEM; and (iii) self-consistency, taking the majority vote over $N = 5$ samples, for Skywork-o1-8B and Marco-o1-7B on GSM8K (Appendix D.11 and D.12).

### 2.2 Causal intervention design

To test the over-generation hypothesis, we imposed brevity constraints on identical problem sets using two prompt conditions: control (unconstrained generation) and brief, a dataset-specific brevity instruction (under 50 words for mathematics, 10 words or fewer for BoolQ, and the answer letter with one sentence for multiple choice; Appendix D.5). Both conditions were applied to seven models (2B–405B) across 115 inverse scaling problems—defined as items where large models ($\geq$70B) underperformed small models ($\leq$9B) in Phase 1—yielding 1,441 paired evaluations (Appendix D.3). Manual error categorisation of 58 failures (29 control, 29 brief) was performed by two independent annotators, with condition labels (control vs brief) hidden from the second annotator to prevent expectation bias. Errors were classified into two categories: premature termination (an initially correct solution path abandoned before a final answer is reached, typically in favour of unfocused elaboration, despite available tokens) and error propagation (faulty logic carried through subsequent reasoning steps). Inter-annotator agreement was high (Cohen's $\kappa = 0.886$; 94.7% raw agreement).

### 2.3 Token efficiency metric

We define token efficiency at generation budget $b$ as accuracy per generated token:

$$\eta(b) = \frac{A(b)}{\bar{T}(b)} \tag{1}$$

where $A(b)$ is accuracy and $\bar{T}(b)$ the mean number of generated tokens at budget $b$. For cross-model correlation analyses we use $\eta_{\text{peak}}$, the value of $\eta(b)$ at the budget where each model attains peak accuracy. The efficiency ratio between two models at matched parameter scale and matched budget is:

$$R(b) = \frac{\eta_{\text{high-}Q}(b)}{\eta_{\text{low-}Q}(b)} \tag{2}$$

For Skywork-o1-8B versus Marco-o1-7B, $R = 5.0$, 10.1 and 18.8 at $b = 1{,}000$, 2,000 and 4,000 tokens. Skywork's generation is budget-invariant (292.8 tokens and 86% accuracy at every budget), whereas Marco's token use grows with budget (961.9 $\rightarrow$ 1,853.8 $\rightarrow$ 3,464.8 tokens) without accuracy gain (56% $\rightarrow$ 54% $\rightarrow$ 54%). At $b = 4{,}000$, $\eta_{\text{Skywork}} = 0.86/292.8 = 0.00294$ and $\eta_{\text{Marco}} =$

$0.54/3{,}464.8 = 0.000156$; the resulting $R = 18.8$ decomposes into an accuracy ratio ($0.86/0.54 \approx 1.6\times$) and a token-usage ratio ($3{,}464.8/292.8 \approx 11.8\times$). We report the range 5.0–18.8× throughout, noting that the upper value depends on the budget ceiling. For self-consistency, $\bar{T}$ is the total number of tokens summed over all samples (Appendix F.4 and F.5).

### 2.4 Mechanistic dimensions

Four observable properties were extracted from reasoning chains generated at the 300-token baseline and maximum-token conditions, without access to model internals. They target three failure modes: quality collapse (vocabulary diversity and repetition rate), regulation failure (token budget utilisation), and verbose decomposition (step decomposition); Section 3.3 reports these as three dimensions.

**Vocabulary diversity.** For a generated response with word sequence $w_1, w_2, \ldots, w_N$, we compute the type–token ratio:

$$D = \frac{|\{w_1, w_2, \ldots, w_N\}|}{N} \tag{3}$$

Diversity change $\Delta D$ is the percentage difference between the maximum-token and 300-token conditions.

Because type–token ratio declines with text length, $\Delta D$ partly reflects output length; it is therefore used comparatively across models evaluated under identical token conditions and interpreted jointly with repetition rate, rather than as an absolute measure of reasoning quality.

**Repetition rate.** Given the set of unique bigrams $B = \{(w_i, w_{i+1})\}$ in a response, the repetition rate is the fraction of bigram types occurring more than once:

$$R_{\text{rep}} = \frac{|\{b \in B : \text{count}(b) > 1\}|}{|B|} \tag{4}$$

**Token budget utilisation.** Self-regulation capacity is measured as the fraction of the allocated token budget consumed:

$$U = \frac{\bar{T}_{\text{generated}}}{T_{\text{limit}}} \tag{5}$$

Models are classified as strongly self-regulating when $U < 0.50$ with ceiling-hit rates below 5%, where ceiling-hit is defined as $T_{\text{generated}} \geq 0.95 \times T_{\text{limit}}$.

**Step decomposition.** Reasoning steps are identified through pattern matching on explicit markers ("Step $n$", numbered sequences, ordinal transitions). Step decomposition efficiency is the mean tokens per reasoning step:

$$S = \frac{\bar{T}_{\text{generated}}}{\bar{n}_{\text{steps}}} \tag{6}$$

### 2.5 Quality score construction

Diversity maintenance and self-regulation exhibited strong correlation ($\rho = +0.88$, $P = 0.002$), indicating a shared underlying mechanism; these were consolidated into a single Quality Control dimension. Step decomposition showed no significant correlation with these components ($r = +0.25$, $P = 0.51$) and was retained as a separate dimension. Each base dimension was normalised to $[0, 1]$ via min–max scaling on the training set ($n = 9$ models), with held-out models ($n = 10$) normalised using training-set bounds.

The composite quality score $Q \in [0, 1]$ integrates these two orthogonal dimensions:

$$Q = 0.7 \times \underbrace{\left(0.5\,\hat{D} + 0.5\,\hat{U}\right)}_{\text{Quality Control}} + 0.3 \times \hat{S} \tag{7}$$

where $\hat{D}$, $\hat{U}$ and $\hat{S}$ denote the min–max-normalised diversity, regulation and decomposition scores respectively (higher values indicate better quality in each dimension: $\hat{D} = 1 - |\Delta D|/100$; $\hat{U} = 1 - U$; $\hat{S} = 1 - (S - S_{\min})/(S_{\max} - S_{\min})$). The 0.7/0.3 weighting reflects the empirical finding that quality maintenance and self-regulation (consolidated as Quality Control) account for more variance in efficiency profiles than step granularity alone. The weighting was selected by silhouette-score maximisation on the training set; across weightings from 0.6/0.4 to 0.8/0.2 the Q–efficiency correlation varies by at most $\Delta r = 0.061$ (Appendix H.2). Q-scores are task-dependent: the same model's score can shift across datasets (Appendix Table 24), reflecting genuine mechanistic differences across task types rather than framework instability. Five interpretive tiers were defined by threshold: Excellent ($Q \geq 0.88$), Good ($Q \geq 0.79$), Moderate ($Q \geq 0.58$), Poor ($Q \geq 0.38$), and Insufficient ($Q < 0.38$). Q-scores are computed on GSM8K. The framework was calibrated on nine models and validated on ten held-out models from five independent families, with the DeepSeek distillation series excluded from calibration to test cross-methodology generalisation Guo et al. (2025); He et al. (2025). Held-out models contributed to neither normalisation bounds, weighting, nor tier thresholds, so held-out tier assignment provides validation independent of the calibration procedure (Appendix G and H).

### 2.6 Statistical analysis

Paired accuracy comparisons across token limits used McNemar's exact test Pembury Smith and Ruxton (2020), with perfect plateaus defined as $\Delta_{\text{acc}} = 0$ ($P = 1.00$). Confidence intervals for proportions used Wilson score intervals. Correlation analyses employed Spearman rank correlation Zar (2005) for ordinal relationships and Pearson correlation Pearson (1896) for continuous associations. Effect sizes were reported as Cohen's $h$ for proportional differences. Multiple comparison corrections applied Bonferroni adjustment where appropriate. Causal intervention effects on GSM8K were assessed with paired $t$-tests; directional consistency across model–dataset conditions with exact binomial sign tests; and single-shot versus self-consistency accuracy with McNemar's exact test. For correlations, the minimum detectable effect at 80% power was computed via the Fisher $z$-transform. Item-level tests pooled across models treat problems as independent and are reported alongside per-model results. All hypothesis tests were two-tailed with $\alpha = 0.05$, except the sign tests, which test a directional hypothesis and are one-sided.

## 3 Results

### 3.1 Near-Universal convergence at 1,000 tokens

Standard models converge to stable accuracy at approximately 1,000 generated tokens regardless of architecture or scale. Token ablation experiments across eight models (1B to 405B parameters) and three reasoning tasks Cobbe et al. (2021a); Clark et al. (2019, 2018) reveal that 21 of 24 model–task conditions exhibit zero accuracy change beyond the 1,000-token boundary ($\Delta = 0.0$ percentage points, $P = 1.00$; Fig. 1A, Appendix Figures 8 and 7). This boundary holds across more than two orders of magnitude in model scale and across mathematically intensive (GSM8K), linguistically demanding (BoolQ) and knowledge-based (ARC-Easy) tasks, and under temperature sampling ($T = 0.6$, 9/9 conditions; Appendix D.11), suggesting a general property of standard models on these benchmarks. The sole exceptions—Llama-3.1-405B conditions—show non-significant trends ($\Delta \leq +3.0$ pp, $P \geq 0.506$) consistent with this model's atypically large capacity ($\geq$87% accuracy at 1,000 tokens across all three tasks; Appendix D.1.1, Table 4).

Among the 21 plateau conditions, 8 represent genuine convergence (improvement then plateau), 12 capacity limitation (flat from the lowest limit) and 1 declines before stabilising (Llama-3.1-8B, GSM8K); the genuine-convergence conditions locate the boundary, while the capacity-limited conditions are consistent with it but do not independently locate it (Appendix Figure 7).

If models had exhausted their reasoning capacity at 1,000 tokens, constraining output length should harm performance. The opposite occurs. Causal intervention experiments—adding a brevity instruction to the prompt versus unconstrained generation (Section 2)—on seven models across five datasets

($n = 1{,}441$ total responses) indicate that constraints *improve* accuracy. On GSM8K inverse-scaling problems Cobbe et al. (2021a), brief constraints raised large-model accuracy from 80.8% to 96.2%, a gain of $+15.4$ percentage points (paired $t_{51} = 2.68$, $P = 0.010$, $n = 52$ per condition; Fig. 1B, Appendix Table 6 and Appendix D.3). The effect generalises: across all datasets, brief constraints reduced the mean performance gap between small and large models from $+38.3$ pp to $+8.2$ pp (78.6% relative reduction, $P = 9.3 \times 10^{-12}$ across 115 matched problems, treating problems as independent; Appendix Table 5). Because these items were selected for inverse scaling, the magnitude may not generalise to typical problems; the random-sample experiment below addresses this. Only BoolQ showed minimal effect ($-0.7$ pp), consistent with its binary response format requiring little elaboration. A complementary experiment on randomly sampled problems (four models, 14B–405B, $n = 100$–$200$ per condition) confirms the effect generalises beyond the inverse-scaling subpopulation, with brief constraints improving accuracy in all eight model–dataset conditions ($+3.0$–$+16.0$ pp; one-sided sign test $P = 0.004$). A length-only condition, with a prompt byte-for-byte identical to control and only the generation cap reduced, improved accuracy in six of eight conditions (directionally consistent but non-significant; one-sided sign test $P = 0.145$), separating the budget effect from prompt framing; the effect also holds under temperature sampling (4/4 conditions; Appendix D.6 and D.11).

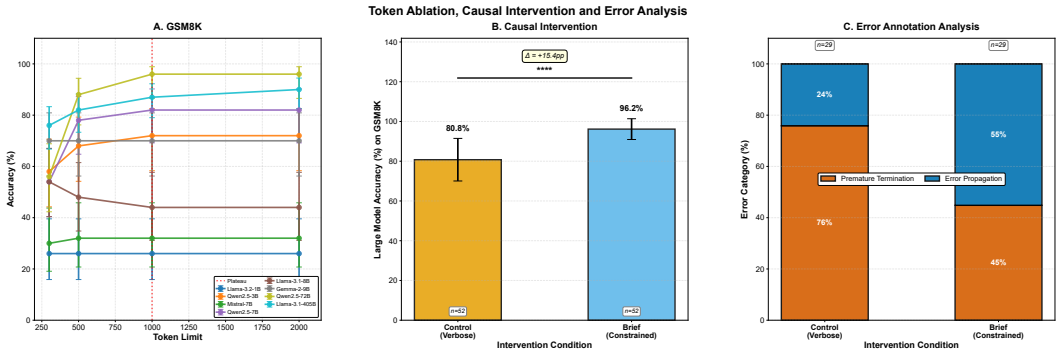



Figure 1: **Token ablation, causal intervention, and error mechanism. (A)** GSM8K accuracy across eight models (1B–405B) at four token limits. Almost all models plateau at 1,000 tokens (red dashed line); BoolQ and ARC-Easy confirm this boundary (Appendix Figure 8). **(B)** Brief constraints improve large-model GSM8K accuracy by $+15.4$ pp (80.8%→96.2%; $P = 0.010$, $n = 52$). **(C)** Blinded error annotation ($n = 58$; $\kappa = 0.886$): premature termination drops from 76% to 45% under constraints, consistent with the gain in panel B.

Blinded two-annotator labelling of 58 incorrect responses (29 per condition; $\kappa = 0.886$, Appendix D.7) identifies the mechanism underlying this effect (Fig. 1C, Appendix Table 9). Under unconstrained generation, 76% of failures involved premature abandonment of correct reasoning despite available tokens, with 24% attributable to error propagation through faulty logic. Constraints reduced premature termination to 45% ($-31$ pp) while error propagation rose to 55% ($+31$ pp)—consistent with a tradeoff in which preventing unfocused elaboration outweighs the cost of occasionally forcing early termination. Because equal numbers of failures were annotated per condition, this analysis characterises the composition of failures rather than their number. Unconstrained responses in this experiment averaged 202 words (brief: 81 words; Appendix D.8), well below any budget considered here. Separately, in the token ablation, models with 2.5-fold differences in baseline verbosity plateau at the same 1,000-token boundary, indicating that additional budget beyond this point largely goes unused rather than being converted into accuracy (Appendix D.8, Appendix Figure 9).

### 3.2 Training quality exceeds parameter scaling

Beyond the aggregate plateau, reasoning-specialised models diverge into three qualitatively distinct efficiency profiles plus a fourth, capacity-limited group (Figure 2, Panel A; Appendix E).

*Early-plateau* models ($n = 3$; 23%) reach near-ceiling accuracy within 1,000 tokens: QWENQWQ-32B(Yang et al., 2024) achieves 92% GSM8K accuracy at 1,000 tokens with no significant subsequent gain ($\Delta = +6$ pp, $P = 0.17$, McNemar's test; Appendix E.5, Appendix Table 15). *Late-blooming* models ($n = 4$; 31%) require extended computation: DEEPSEEK-R1-685B(DeepSeek-AI, 2025) gains $+56$ pp from 1,000 to 4,000 tokens on GSM8K ($P < 0.001$), while QWEN3-32B(Yang et al., 2025) gains $+24$ pp between 1,000 and 2,000 tokens ($P = 0.001$). Parameter scale does not predict

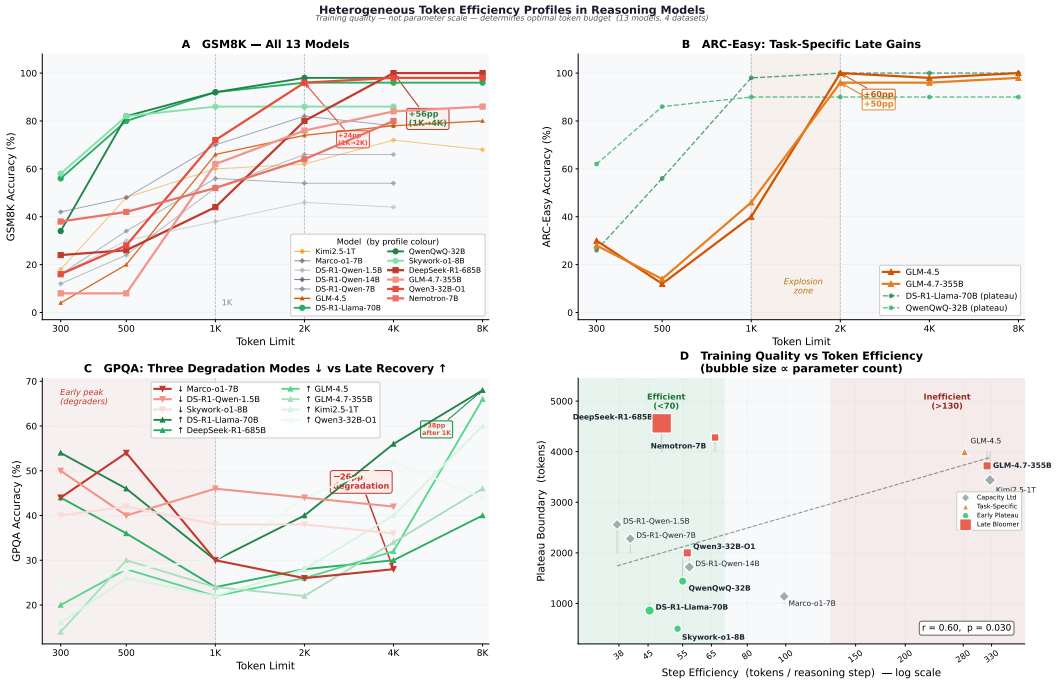



Figure 2: **Heterogeneous token efficiency profiles in reasoning models. (A)** GSM8K trajectories across six token limits. Colour encodes profile: red (late bloomer), green (early plateau), orange (task-specific), grey (capacity-limited). **(B)** ARC-Easy late gains: GLM-4.5 and GLM-4.7 improve by $+60$ pp and $+50$ pp at the 1 000-to-2 000-token transition ($P < 0.001$). **(C)** GPQA degradation modes: over-generation (Marco-o1-7B, $-26$ pp), capacity overflow (DS-R1-1.5B), and early-peak decay (Skywork-o1-8B). **(D)** Step efficiency versus plateau boundary ($r = +0.60$, $P = 0.030$); bubble size proportional to parameter count.

this behaviour: NEMOTRON-7B(Ahmad et al., 2025) (7B) and DEEPSEEK-R1-685B (685B) share this profile despite a 98-fold size difference (Appendix Table 13 and Table 15). A single *task-specific* model (GLM-4.5(Team et al., 2025)) shows discontinuous gains confined to particular datasets: $+60$ pp on ARC-Easy(Clark et al., 2018) at the 1,000-to-2,000-token transition ($P < 0.001$; Figure 2, Panel B), while GSM8K saturates at 80%. The remaining five models (38%) are *capacity-limited*, showing no token-driven improvement beyond 2,000 tokens despite peak accuracies spanning 46% (DS-R1-1.5B) to 82% (DS-R1-14B)(DeepSeek-AI, 2025) (Appendix E.2).

Across all reasoning models, step efficiency (tokens per reasoning step) correlates with plateau boundary ($r = +0.60$, $P = 0.030$; Figure 2, Panel D), indicating that models producing more concise reasoning steps plateau earlier, independent of parameter count.

**A 5.0–18.8-fold efficiency advantage from training quality** Direct comparison at matched parameter scales establishes training methodology as the primary efficiency determinant. We quantify training quality as a composite score $Q \in [0, 1]$ integrating vocabulary-diversity maintenance and budget self-regulation (Quality Control) with step decomposition efficiency, computed on GSM8K (Section 2; Appendix G and H, Table 19). Within the 7–8B range, SKYWORK-O1-8B ($Q = 0.9867$)(He et al., 2025) achieves 86% GSM8K accuracy at 1,000 tokens using 292.8 average tokens, while MARCO-O1-7B ($Q = 0.2428$)(Zhao et al., 2024) peaks at 56% at 1,000 tokens yet consumes 3,464.8 average tokens at the 4,000-token limit (54%, degraded) (Appendix Table 19 and Table 17). At a matched 4,000-token budget this is an 18.8-fold difference in accuracy per token consumed (0.00294 vs. 0.000156), falling to 10.1-fold at 2,000 tokens and 5.0-fold at 1,000 tokens (Section 2; Appendix F.4). This ratio comprises a 1.6× accuracy component (86%/54%) and an 11.8× self-regulation component (3,465/293 tokens). On ARC-Easy and BoolQ, where both models converge at moderate token budgets, Marco-o1-7B achieves higher peak accuracy, indicating the efficiency differential is task-dependent and most pronounced on mathematical reasoning. This training-quality effect exceeds the efficiency gains from ten-fold parameter scaling: Skywork-o1-8B's token efficiency

($\eta = 0.00294$) is 1.8 times that of DEEPSEEK-R1-70B ($\eta_{\text{peak}} = 0.00162$), despite the latter's higher accuracy (96% vs. 86%); this is a cross-family comparison rather than a controlled scaling experiment. The complete efficiency gradient across all five quality tiers—from Tier 1 ($Q = 0.987$, $\eta = 0.00294$) to Tier 5 ($Q = 0.082$, $\eta = 0.000235$)—is documented in Appendix Tables 17 and 19, confirming a strong positive relationship between training quality and token efficiency across all models (Pearson $r = +0.796$, $P = 0.0001$, 95% CI [0.52, 0.92], $r^2 = 0.634$, $n = 18$; Appendix H.2) across models spanning 1.5B–685B parameters. Because the weighting was selected on the calibration set and utilisation shares token counts with $\eta$, this correlation is partly internal (excluding utilisation still gives $r = +0.826$ on the 13 original models); held-out tier assignment (10/10 models; Section 2) provides validation independent of calibration.

The DeepSeek-R1 distillation series links Q to a documented training gradient: five models distilled from the same teacher with the same 800k samples Guo et al. (2025) into progressively smaller students (70B→1.5B) yield monotonically decreasing Q-scores ($0.873 \rightarrow 0.818 \rightarrow 0.684 \rightarrow 0.359 \rightarrow 0.082$; Appendix H.6 and H.6).

Task complexity further modulates token requirements. BoolQ (Clark et al., 2019) is largely insensitive to budget: 9 of 10 eligible models show no meaningful gain beyond 1,000 tokens (Appendix E.2, Figure 12). ARC-Easy produces the most discontinuous improvements, with step-function gains exceeding $+50$ pp at specific thresholds. No single token budget satisfies all model–task combinations, necessitating empirical profiling for adaptive allocation.

### 3.3 Mechanistic basis of efficiency differences

For the 13 fully profiled models, three properties measured directly from generated outputs, without access to model internals—vocabulary diversity, self-regulation and step decomposition—explain *why* these profiles arise (Figure 3; Appendix I, Table 27).

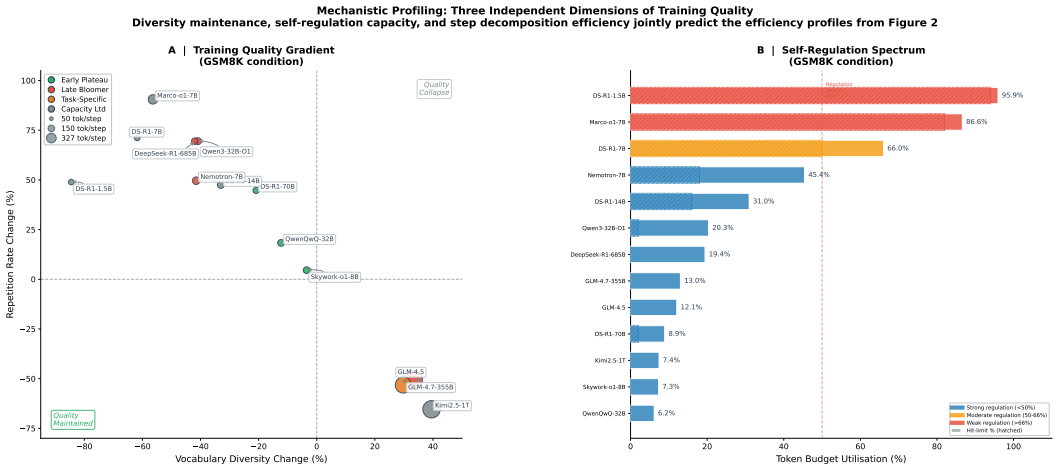



Figure 3: **Two mechanistic dimensions of training quality (GSM8K condition). (A)** Vocabulary diversity versus repetition rate change from 300 to maximum tokens. Lower-right = quality maintained; upper-left = quality collapse. **(B)** Token budget utilisation across reasoning models (6.2%–95.9%). Hatching = ceiling-hit rate; red dashed line = regulation failure boundary. Step decomposition in Appendix Figure 19.

**Quality maintenance.** From 300 tokens to the model maximum on GSM8K (Figure 3A), three models (KIMI2.5 (Team et al., 2026), GLM-4.7 (Team et al., 2025), GLM-4.5 (Team et al., 2025)) increase diversity ($+30\%$ to $+40\%$) while reducing repetition ($-50\%$ to $-65\%$); the remaining ten models show the opposite pattern, with the most severe cases (DS-R1-1.5B (DeepSeek-AI, 2025): $-84.5\%$ diversity, $+48.9\%$ repetition; MARCO-O1-7B (Zhao et al., 2024): $-56.4\%$, $+90.5\%$) corresponding to capacity-limited profiles. Quality maintenance is task-dependent—GLM-4.5 and GLM-4.7 improve on GSM8K but collapse on BoolQ ($-68.5\%$ and $-60.1\%$ diversity; Appendix Table 27).

**Self-regulation.** Token budget utilisation spans an order of magnitude (Figure 3B): from 6.2% (QWENQWQ-32B (Yang et al., 2024)) to 95.9% (DS-R1-1.5B (DeepSeek-AI, 2025)) on GSM8K. Ten models use <50% of their allocation, eight of them with ceiling-hit rates ≤2% (Nemotron-7B: 18%; DS-R1-14B: 16%), indicating largely voluntary termination; three models (DS-R1-1.5B: 94%, MARCO-O1-7B: 82%, DS-R1-7B: 50% ceiling-hit) exhaust their budgets consistently (Appendix Table 27).

**Step decomposition.** Mean tokens per reasoning step separates the cohort into two non-overlapping clusters (Appendix Figure 19): a compact group (37.6–99.1 tokens/step, $n = 10$) and a verbose group (282.2–327 tokens/step, $n = 3$), with no model between 100 and 282 tokens/step. This structural gap is consistent with the ARC-Easy discontinuity (Section 3.2): GLM-4.5 and GLM-4.7 gain +60 and +50 pp at 2,000 tokens, when the budget first accommodates a complete reasoning step (Appendix Table 17); truncation of the final answer below 2,000 tokens would produce the same pattern and is not excluded. KIMI2.5 (Team et al., 2026) illustrates the interaction: despite the largest diversity gain (+39.5%) and strong self-regulation (7.4% usage), its verbose decomposition (327 tokens/step) caps GSM8K accuracy at 72% (Appendix Table 17 and Table 27).

Self-regulation and step decomposition are not significantly associated ($r = -0.357$, $P = 0.231$, $n = 13$, limited power; Appendix Table 28), so we treat budget exhaustion and verbose reasoning as distinct failure modes.

### 3.4 Graduate-level quality collapse

Twelve of the thirteen profiled reasoning models (92%) exhibit quality collapse on graduate-level questions (GPQA) (Rein et al., 2024)—vocabulary diversity falling by 18–73% as the budget grows, versus 3.7% for the exception—despite adequate arithmetic performance, a failure mode invisible on standard benchmarks (Figure 2, Panel C).

**Paradoxical efficiency signature.** Compact GSM8K step decomposition (37.6–99.1 tokens/step) does not protect against GPQA failure: MARCO-O1-7B (Zhao et al., 2024) (99.1 tokens/step on GSM8K) peaks at 54% GPQA accuracy at 500 tokens then degrades to 28% at 4,000 tokens ($\Delta = -24$ pp at the sharpest step, $P = 0.015$; Appendix Table 17).

Even DEEPSEEK-R1-685B (DeepSeek-AI, 2025), which maintains the best vocabulary diversity on GPQA (−3.7% change vs. −18.1% to −72.9% for others; Appendix Table 27), achieves only 40% peak GPQA accuracy—demonstrating that diversity maintenance alone is insufficient for graduate-level reasoning.[1]

**Self-regulation breakdown.** Greater budget use on harder problems would be expected of a well-calibrated model, but here it is not matched by accuracy gains: MARCO-O1-7B rises from 86.6% usage on GSM8K to 98.2% on GPQA (98% ceiling-hit) while its accuracy falls from 54% to 28%, and KIMI2.5 rises from 7.4% to 70.9% (utilisation is relative to each model's own cap). Among seven models exhibiting strong regulation on GSM8K (usage <20%), five (71%) consume ≥50% of their budget on GPQA, indicating that graduate-level tasks systematically exhaust token allocations. The two exceptions, QWENQWQ-32B (19.8%) and SKYWORK-O1-8B (42.9%) (He et al., 2025), were both trained with reinforcement learning emphasising generation quality (Appendix Table 27).

**Benchmark validity implications.** GSM8K and GPQA peak accuracies show a near-zero, non-significant correlation ($r = -0.027$, $P = 0.915$, $n = 18$; power sufficient to detect $r \geq 0.6$ but not to exclude moderate correlations; Appendix E.4.1 and Table 14). Among eleven models achieving ≥80% GSM8K accuracy (Appendix Tables 13 and 14), GPQA peak accuracy spans 40–74% (mean ≈56%), on average ∼34 pp below their GSM8K accuracy; because GPQA is four-choice (25% chance) and GSM8K open-ended, this gap is not a like-for-like difficulty measure. This cautions against characterising reasoning efficiency exclusively on grade-school benchmarks.

[1] DeepSeek-R1-685B does not comply with token constraints below 1,000 tokens on GPQA; all claims use compliant conditions only (Appendix D.10.5).

## 4 Discussion and Limitations

Across the models and 67,000 evaluations examined here, training quality, rather than parameter count or inference token budget, is the strongest determinant of reasoning efficiency we observe. This complements scaling laws Kaplan et al. (2020); Hoffmann et al. (2022), which describe how much capability training acquires; the Densing Law Xiao et al. (2025), which describes how quickly that frontier advances; and evidence that RLVR improves sampling efficiency without expanding reasoning boundaries Yue et al. (2026). The present work characterises the inference-side complement: how effectively acquired capability is *expressed* at deployment, evidenced by the 5.0–18.8-fold efficiency difference at matched parameter scales. Within the evaluated models and benchmarks, efficiency boundaries are near-universal for standard models (approximately 1,000 tokens), strongly associated with training quality among reasoning models (Appendix H.6), and measurable from observable chain properties without access to model internals.

These findings reframe test-time compute allocation as a quality-dependent decision rather than a uniform scaling problem. Early-plateau models need only 1,000 tokens to reach their ceiling, while capacity-limited models consume roughly $4\times$ more tokens to reach lower ceilings (Appendix Table 17). Training interventions targeting *termination calibration* may yield greater returns than approaches emphasising extended reasoning Silver et al. (2016); Madaan et al. (2023); Cobbe et al. (2021a): additional tokens are productive only when models maintain quality under extended generation.

Because GSM8K accuracy does not predict GPQA accuracy ($r = -0.027$) and 12 of 13 profiled models collapse on GPQA, efficiency characterised on grade-school tasks should not be assumed to transfer; the mechanistic dimensions introduced here offer complementary evaluation axes capturing *how* models reason, not only *whether* they answer correctly.

Future work should manipulate termination versus correctness rewards during reinforcement learning to establish causal training pathways, develop adaptive allocation that monitors diversity collapse and budget use in real time, and force earlier or later termination to distinguish an autoregressive information bottleneck from natural stopping behaviour as the source of the 1,000-token convergence.

Several limitations constrain these findings. The diagnostic Q-score correctly assigns all ten held-out models from five families to their tier, but ten models still leave substantial uncertainty (95% CI [69.2%, 100%]; Appendix H); its weighting was selected on the calibration set, so the Q–efficiency correlation is partly internal, and it characterises behaviour on the calibration task rather than robustness under the shift to graduate-level reasoning, where even the top-ranked model degrades (Section 3.4). The framework is calibrated on GSM8K, with cross-dataset Q-score shifts of up to 0.74 points (Appendix Table 24), indicating that tier thresholds require task-specific recalibration. The mechanistic dimensions are observational; the causal training experiments proposed above remain necessary to confirm directional relationships. Causal interventions targeted inverse scaling problems (7.7% of conditions); a complementary random-sample experiment confirms the effect generalises (Appendix D.6), though individual conditions remain underpowered (3 of 24 McNemar tests significant at $P < 0.05$). Sample sizes ($n = 50$ per condition) limit detection power to effects $\geq 18$ pp. The primary experiments use greedy decoding to isolate training-quality effects from sampling stochasticity; targeted replications under temperature sampling confirm that the convergence boundary (three models) and causal intervention effect (two models) persist (Appendix D.11), and self-consistency on the headline matched-scale pair narrows but does not close the efficiency gap ($5.3\times \rightarrow 3.3\times$; Appendix D.12). Validation of best-of-$N$ and verifier-guided search across the full cohort remains for future work. The study examines 19 open-source reasoning models released between November 2024 and February 2026; validation on proprietary systems such as o1 Jaech et al. (2024) and Gemini Comanici et al. (2025) would further establish generalisability. For standard models that self-terminate well below 1,000 tokens, the absence of change at larger budgets reflects natural stopping rather than a tested saturation of reasoning; the 8 genuine-convergence conditions carry the boundary claim. The upper value of the efficiency ratio ($18.8\times$) depends on the budget ceiling and derives from a single matched pair. GPQA estimates rest on a 50-item stratified subsample, peak accuracy taken over multiple budgets is optimistically biased, and GSM8K scores near ceiling attenuate the cross-benchmark correlation. Finally, the error analysis ($n = 58$, $\kappa = 0.886$) covers only two error categories on a single dataset.

## Acknowledgments and Disclosure of Funding

This research did not receive any specific grant or funding from any public, commercial, or not-for-profit organisation during its conduct. The author declares no competing interests. The author thanks the anonymous reviewers and the area chair for their constructive feedback, which substantially improved this work.

## References

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric P. Xing, Sham M. Kakade, and Hanlin Zhang. EvoLM: In search of lost language model training dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=B6bE2GC71a.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *Nature Machine Intelligence*, pages 1–11, 2025.

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

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. An empirical analysis of compute-optimal inference for problem-solving with language models. 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021a.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Matilda QR Pembury Smith and Graeme D Ruxton. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74(11):133, 2020.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.

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, 2021b. URL https://arxiv.org/abs/2110.14168.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

An Yang, Anfeng Li, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. Opencodereasoning: Advancing data distillation for competitive coding. 2025. URL https://arxiv.org/abs/2504.01943.

GLM Team, Aohan Zeng, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions, 2024. URL https://arxiv.org/abs/2411.14405.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, and Others. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=4OsgYD7em5.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

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*, volume 35, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do not think that much for 2+3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

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 preprint arXiv:2501.19393*, 2025.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vsgq2ldr4K.

Jian Huang et al. Hapo: History-aware efficient large reasoning models via bi-level policy gradient optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

others Liu. Laser-d: Difficulty-aware step reward for mathematical reasoning. *arXiv preprint*, 2025.

others Luo. O1-pruner: Length-harmonizing thinking llms via sequential kv cache pruning. *arXiv preprint*, 2025.

others Yuan. Short-rl: Reinforcement learning for efficient and concise reasoning. *arXiv preprint*, 2025.

others Song. S-grpo: Early exit via synchronized group relative policy optimization. *arXiv preprint*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

# Appendix

## Appendix Contents


**A Extended Related Work** **16**

**B The Step-Accuracy Paradox** **17**

B.1 Discovery of Inverse Correlation . . . . . 17

B.2 Task-Dependent Paradox Strength . . . . . 17

B.3 Persistence Across Model Scales . . . . . 17

B.4 Implications for Test-Time Compute Scaling . . . . . 18

**C Details Results for The Step-Accuracy Paradox Section** **18**

C.1 Step-Accuracy Paradox Statistics . . . . . 18

C.2 Reasoning Length Distributions by Model Family . . . . . 19

C.3 Per-Model Correlation Statistics. . . . . 20

C.4 Question and Model Characteristic Correlations with Accuracy . . . . . 22

C.5 Correlation Matrix of Question and Model Characteristics . . . . . 22

**D Near-Universal convergence at 1,000 tokens** **24**

D.1 Complete Token Ablation Results Of Standard Models . . . . . 24

D.2 Token Ablation: BoolQ and ARC-Easy . . . . . 26

D.3 Complete causal intervention results . . . . . 27

D.4 Statistical validation of causal intervention effects . . . . . 28

D.5 Intervention prompts used in causal experiments . . . . . 29

D.6 Random-sample validation and prompt-confound separation . . . . . 29

D.7 Manual error categorization by intervention condition . . . . . 31

D.8 Token Utilization Pattern of Standard Models . . . . . 31

D.9 Model-specific error patterns . . . . . 32

D.10 Statistical Methods and Additional Notes . . . . . 33

D.11 Temperature sampling replication . . . . . 35

D.12 Self-consistency and multi-sample strategies . . . . . 36

**E Training quality exceeds parameter scaling** **36**

E.1 Profile Assignment Criteria . . . . . 36

E.2 Complete Accuracy by Token Limit . . . . . 37

E.3 Individual Model Trajectory Figures . . . . . 38

E.4 GSM8K-GPQA Correlation Analysis . . . . . 40

E.5 Complete Pairwise Statistical Analysis . . . . . 41

E.6 Token Utilisation at the 300-Token Limit . . . . . 45

E.7 Statistical Notes . . . . . 46

**F Training Quality Efficiency Improvement at Matched Parameter Scales** **47**
F.1 Overview . . . 47
F.2 Calculation Methodology . . . 47
F.3 Experimental Data . . . 47
F.4 Efficiency Improvement Calculation . . . 48
F.5 Interpretation . . . 48
F.6 Robustness Validation . . . 48

**G A Diagnostic Framework for Training Quality Assessment** **54**
G.1 Two-Dimensional Quality Score . . . 54
G.2 Quality Tiers and Model Distribution . . . 55
G.3 Validation Results . . . 55

**H Diagnostic Framework Extended Analysis** **56**
H.1 Complete Quality Scores and Validation Results . . . 56
H.2 Q-Score vs Token Efficiency Correlation . . . 58
H.3 Cross-Dataset Generalization Analysis . . . 59
H.4 Statistical Validation . . . 60
H.5 Methodological Details . . . 60
H.6 Distillation gradient mechanistic profile . . . 61

**I Mechanistic basis of efficiency differences** **62**
I.1 Complete Mechanistic Dimensions Across Datasets . . . 62
I.2 Cross-Dataset Quality Maintenance Profiles . . . 64
I.3 GPQA as a Quality Degradation Stress Test . . . 64
I.4 Step Decomposition Efficiency . . . 65

# A Extended Related Work

The present study builds on several lines of prior work.

**Chain-of-thought and test-time compute.** Chain-of-thought prompting Wei et al. (2022) established that explicit reasoning steps improve accuracy, motivating the test-time compute paradigm. Self-consistency decoding Wang et al. (2023) demonstrated that sampling multiple reasoning paths can substitute for longer individual chains, while best-of-$n$ sampling and reward-guided search Snell et al. (2024); Lightman et al. (2023) introduced step-level quality signals as alternatives to unconstrained generation. The present work complements these approaches by characterising when additional tokens within a *single* chain transition from beneficial to harmful.

**Overthinking and verbosity in reasoning models.** Recent work has identified "overthinking" in reasoning models Chen et al. (2024), where extended generation degrades performance on simple problems, and budget-forcing experiments Muennighoff et al. (2025) have shown that constraining generation length can improve accuracy. Our causal intervention experiments (Section 3.1) extend these findings across 50 models and six benchmarks, establishing the over-generation mechanism as near-universal rather than model-specific, and identifying the 1,000-token convergence boundary absent from prior work.

**Scaling laws and training efficiency.** Neural scaling laws Kaplan et al. (2020); Hoffmann et al. (2022) characterise how performance improves with parameters and data during training. The densing law Xiao et al. (2025) extends this to temporal efficiency, showing capability density doubles every 3.5 months. Recent systematic analysis of training-stage decisions Qi et al. (2025) demonstrates that pre-training composition and post-training methodology produce significant capability differences. The present work characterises the inference-side complement: how training quality determines the efficiency with which learned capability is expressed at deployment.

**Capability boundaries of post-training methods.** Recent work Yue et al. (2026) demonstrates that reinforcement learning with verifiable rewards (RLVR) improves single-sample accuracy (pass@1) but does not expand the set of solvable problems: base models achieve equivalent or superior pass@$k$ at $k \geq 128$ across all tested benchmarks and model families. Perplexity analysis confirms that RLVR-generated reasoning paths already exist within the base model's sampling distribution. In contrast, distillation from a stronger teacher genuinely expands the reasoning boundary. These findings complement the present work by providing a training-side explanation for why training methodology produces the efficiency differences we observe at inference time: RLVR redistributes sampling probability toward correct paths (improving token efficiency for well-trained models) without creating new capability, while distillation can introduce reasoning patterns absent from the base model's prior.

Concurrent work by Karan and Du (2026) provides inference-time evidence for the same distribution-sharpening account. They demonstrate that sampling from the power distribution $p^\alpha$ of a base model—approximated via Metropolis-Hastings MCMC at optimal $\alpha = 4.0$—achieves single-shot accuracy comparable to GRPO on MATH500 (74.8% vs. 78.5%) and outperforms it on out-of-domain tasks, without training, dataset, or verifier access. Their Proposition 3, which formally characterises how low-temperature sampling upweights tokens with many low-likelihood future completions over tokens with singular high-likelihood futures, offers a theoretical account consistent with the over-generation failure we document in Corollary 3. The diversity collapse they document for GRPO—responses concentrated at the highest likelihood peak with pass@$k$ tapering for $k > 1$ (their Figures 4–5)—independently confirms the vocabulary diversity collapse we characterise mechanistically in Section 3.3. Their method naturally produces responses averaging 679 tokens on MATH500, consistent with our 1,000-token convergence boundary. Applying our efficiency metric to their *computational* cost—using their own reported $8.84\times$ token overhead (Appendix A.3; $N_{\text{MCMC}} = 10$, $T = 679$, $B = 192$)—yields $\eta_{\text{compute}} = 0.000125$, placing training-free power sampling below every trained model in our cohort; their GRPO baseline achieves $\eta = 0.00117$ ($9.39\times$ advantage), establishing that training quality determines how cheaply the inference-time accuracy ceiling is accessed.

**Adaptive inference and length-efficient reasoning.** Recent work proposes training interventions to reduce reasoning verbosity. HAPO Huang et al. (2026) introduces a history-aware length penalty during reinforcement learning to discourage unnecessary elaboration. LASER-D Liu (2025) applies difficulty-aware step rewards, allocating more reasoning steps to harder problems. O1-Pruner Luo (2025) uses a pre-sample baseline with RL to prune redundant reasoning. Short-RL Yuan (2025) employs a SlackBand tolerance mechanism that permits concise solutions without penalising correctness. S-GRPO Song (2025) combines early exit with a decaying reward signal to incentivise shorter chains. These methods propose specific algorithmic interventions; the present work is complementary, providing the diagnostic framework (Q-score, mechanistic dimensions, over-generation analysis) that characterises the problem these methods aim to solve and can serve as an evaluation tool for comparing their outcomes.

**Reasoning evaluation and benchmarks.** GSM8K Cobbe et al. (2021a) remains the dominant reasoning benchmark, yet concerns about benchmark saturation have motivated harder evaluations including GPQA Rein et al. (2024) and MATH. Our finding that GSM8K performance fails to predict GPQA behaviour ($r = -0.027$, $P = 0.915$, $n = 18$) provides quantitative evidence for these concerns, demonstrating that efficiency characterised on arithmetic tasks does not reliably transfer to graduate-level reasoning.

# B The Step-Accuracy Paradox

To test the assumption that extended reasoning monotonically improves performance, this study evaluated 31 standard language models (0.5B to 405B parameters) across five reasoning benchmarks (GSM8K, BoolQ, ARC-Easy, CommonsenseQA, MMLU-STEM) with 300 samples each. Model families included Llama, Qwen, Mistral, Gemma, Phi, Yi, GPT, StableLM, DeepSeek, Kimi and Gemini (11 families) spanning diverse scales. For each evaluation, the number of reasoning steps in generated traces was extracted and correlated with final answer accuracy.

## B.1 Discovery of Inverse Correlation

The analysis reveals negative correlation between reasoning steps and accuracy (Spearman $\rho = -0.093$, $p < 10^{-25}$, $n = 44{,}106$; Figure 4A). This step-accuracy paradox contradicts the assumption underlying test-time compute scaling: problems requiring longer reasoning chains show systematically lower accuracy. The effect persists in three of four model-scale bands (Spearman $\rho \approx -0.08$ for tiny, small, and giant models; Table 1), indicating a fundamental pattern rather than artifact of specific model properties.

## B.2 Task-Dependent Paradox Strength

The paradox manifests with task-dependent strength (Figure 4A-B). Mathematical reasoning exhibits the strongest effect: GSM8K shows pronounced negative correlation ($\rho = -0.203$, $p < 10^{-78}$, strong effect), robust across all 31 models. Science knowledge tasks (MMLU-STEM) demonstrate negligible correlation ($\rho = -0.008$, $p = 0.462$), while commonsense reasoning (CommonsenseQA) shows weak negative correlation ($\rho = -0.024$, $p = 0.029$). Binary questions (BoolQ) exhibit no significant relationship ($\rho = -0.012$, $p = 0.257$), consistent with minimal multi-step reasoning requirements.

Notably, retrieval-focused science questions (ARC-Easy) demonstrate positive correlation ($\rho = +0.087$, $p < 10^{-15}$, moderate effect), the sole exception to the paradox pattern. This reversal suggests that multi-hop retrieval tasks benefit from extended reasoning, whereas computational tasks (GSM8K) show quality degradation with verbosity. Effect strength varies from strong (GSM8K: $|\rho| = 0.203$) to negligible (MMLU-STEM, BoolQ: $|\rho| < 0.015$), with only mathematical reasoning consistently showing pronounced paradox across all models (Figure 4B; Appendix Table 1).

## B.3 Persistence Across Model Scales

The paradox persists across model scales, though magnitudes vary (Figure 4C). Small models (2-10B) show negative correlation ($\rho = -0.081$, $p < 10^{-33}$), comparable to tiny models (<2B: $\rho = -0.075$, $p < 10^{-8}$) and giant models (100B+: $\rho = -0.078$, $p = 0.003$). Large models (10-100B) exhibit weak positive correlation ($\rho = +0.028$, $p = 0.003$) (Appendix Table 2), suggesting improved

reasoning utilization at this scale, though effect magnitude remains small. The consistent negative correlation across most scales indicates the paradox is not resolved by increased parameter count alone.

### B.4 Implications for Test-Time Compute Scaling

These findings challenge assumptions underlying current test-time compute scaling paradigms. For the most common reasoning benchmark (GSM8K), extended reasoning correlates strongly and negatively with accuracy, suggesting that verbose reasoning may indicate quality degradation rather

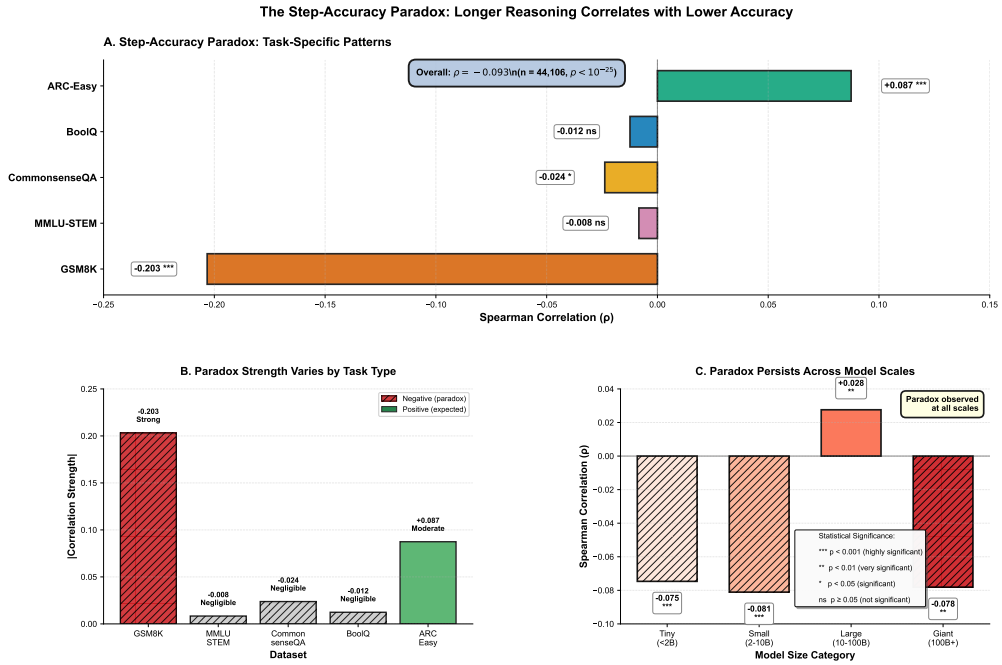



Figure 4: **The Step-Accuracy Paradox Across Tasks and Model Scales. (A)** Correlation between reasoning steps and accuracy varies by task type (44,106 evaluations, 31 models). Mathematical reasoning (GSM8K) shows strong negative correlation ($\rho = -0.203^{***}$), while retrieval tasks (ARC-Easy) show moderate positive correlation ($\rho = +0.087^{***}$). Overall correlation is negative ($\rho = -0.093$, $p < 10^{-25}$). **(B)** Effect strength comparison. GSM8K exhibits strong paradox; MMLU-STEM and BoolQ show negligible effects. Hatched bars: negative correlation; solid bars: positive correlation. **(C)** Paradox persists across model scales (0.5B–405B parameters). Most size categories show negative correlations ($\rho \approx -0.08$). Significance: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, ns $p \geq 0.05$.

than thorough problem-solving. The task-dependent patterns—positive correlation for retrieval (ARC-Easy), strong negative for computation (GSM8K)—indicate that not all reasoning is equal: productive multi-step reasoning differs fundamentally from counterproductive verbosity.

The paradox's persistence across 31 models establishes baseline efficiency boundaries for standard language models. While algorithmic interventions (Monte Carlo tree search, self-verification) may overcome these limitations, the observed patterns motivate systematic investigation of token budget constraints and training quality effects examined in subsequent sections.

## C Details Results for The Step-Accuracy Paradox Section

### C.1 Step-Accuracy Paradox Statistics

The overall negative correlation in Table 1 ($\rho = -0.093$) establishes the step-accuracy paradox as a moderate-strength effect across all evaluated conditions. Task decomposition reveals substantial heterogeneity: mathematical reasoning tasks (GSM8K) exhibit strong paradox ($\rho = -0.203$, effect strength $2.2\times$ overall magnitude), while science knowledge tasks (MMLU-STEM, BoolQ) show

negligible correlations ($|\rho| < 0.015$). Retrieval-focused tasks (ARC-Easy) demonstrate the sole reversal pattern with moderate positive correlation ($\rho = +0.087$), suggesting multi-hop fact retrieval benefits from extended reasoning unlike computational tasks. Model size analysis indicates paradox persistence across scales: tiny, small, and giant models all exhibit negative correlations ($\rho \approx -0.08$), with only large models (10–100B parameters) showing weak positive correlation ($\rho = +0.028$). This size-dependent variation suggests improved reasoning utilization emerges between 10B and 100B parameters but remains modest in magnitude. The consistency of negative correlations across 44,106 evaluations spanning diverse architectures and model families

Table 1: **Complete Step-Accuracy Paradox Statistics.** Spearman rank correlations between reasoning steps and accuracy across 31 standard models (0.5B–405B parameters) on five benchmarks. Overall correlation ($\rho = -0.093$, $p < 10^{-25}$, $n = 44{,}106$) shows significant negative relationship. Task-specific patterns vary from strong negative (GSM8K: $\rho = -0.203$***) to moderate positive (ARC-Easy: $\rho = +0.087$***). Effect strength: strong ($|\rho| \geq 0.15$), moderate ($0.08 \leq |\rho| < 0.15$), weak ($0.03 \leq |\rho| < 0.08$), negligible ($|\rho| < 0.03$). Significance: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns $p \geq 0.05$.

| **Category** | **Subcategory** | **Spearman $\rho$** | **$p$-value** | **Effect** | **Pattern** | $n$ |
|---|---|---|---|---|---|---|
| Overall | All datasets | $-0.093$ | $< 10^{-25}$ | Moderate | Negative | 44,106 |
| By Dataset | GSM8K | $-0.203$ | $< 10^{-78}$ | Strong | Negative | $\sim$8,800 |
| | MMLU-STEM | $-0.008$ | 0.462 | Negligible | Negative | $\sim$8,800 |
| | CommonsenseQA | $-0.024$ | 0.030 | Negligible | Negative | $\sim$8,800 |
| | BoolQ | $-0.012$ | 0.257 | Negligible | Negative | $\sim$8,800 |
| | ARC-Easy | $+0.087$ | $< 10^{-15}$ | Moderate | Positive | $\sim$8,800 |
| By Size | Tiny (<2B) | $-0.075$ | $< 10^{-8}$ | Weak | Negative | varies |
| | Small (2–10B) | $-0.081$ | $< 10^{-33}$ | Moderate | Negative | varies |
| | Large (10–100B) | $+0.028$ | 0.003 | Negligible | Positive | varies |
| | Giant (100B+) | $-0.078$ | 0.003 | Weak | Negative | varies |

establishes the paradox as a fundamental efficiency constraint in current language model reasoning rather than an artifact of specific training procedures or architectural choices.

### C.2 Reasoning Length Distributions by Model Family

In Figure 5 reasoning length distributions reveal systematic family-level patterns in verbosity independent of task difficulty, providing evidence for training methodology effects on reasoning chain length. The 2.7× range in mean steps (GPT: 2.0 to Yi: 5.3) across similarly-capable models indicates verbosity is a learned characteristic rather than an emergent property of reasoning competence. Within-family consistency—evidenced by tight distributions for Qwen, Phi, and Gemma families—suggests organizational training procedures strongly shape reasoning style, with individual model variations primarily reflecting scale effects rather than fundamental approach changes. The Llama family's exceptional distribution width (coefficient of variation ≈2.4, highest among evaluated families) demonstrates that parameter scaling within a consistent architecture induces systematic verbosity increases, with 405B models generating substantially longer chains than 1B models on identical problems. This scale-dependent verbosity amplification, combined with the step-accuracy paradox, suggests larger models may suffer accuracy penalties from over-generation despite superior base capabilities. DeepSeek's bimodal distribution uniquely identifies two distinct reasoning regimes within a single family: one group of models clusters at shorter lengths (mode ≈4 steps) while another favors extended chains (mode ≈20 steps), indicating that models within a single family can adopt markedly different verbosity regimes. The strong family-level clustering despite heterogeneous task difficulties (mathematical GSM8K vs. simple BoolQ comprehension) confirms that models generate reasoning chains of characteristic length regardless of problem complexity, supporting the interpretation that verbosity reflects model-specific tendencies rather than adaptive responses to question difficulty. This finding strengthens the paradox interpretation by demonstrating extended reasoning emerges from model characteristics rather than genuine problem requirements.

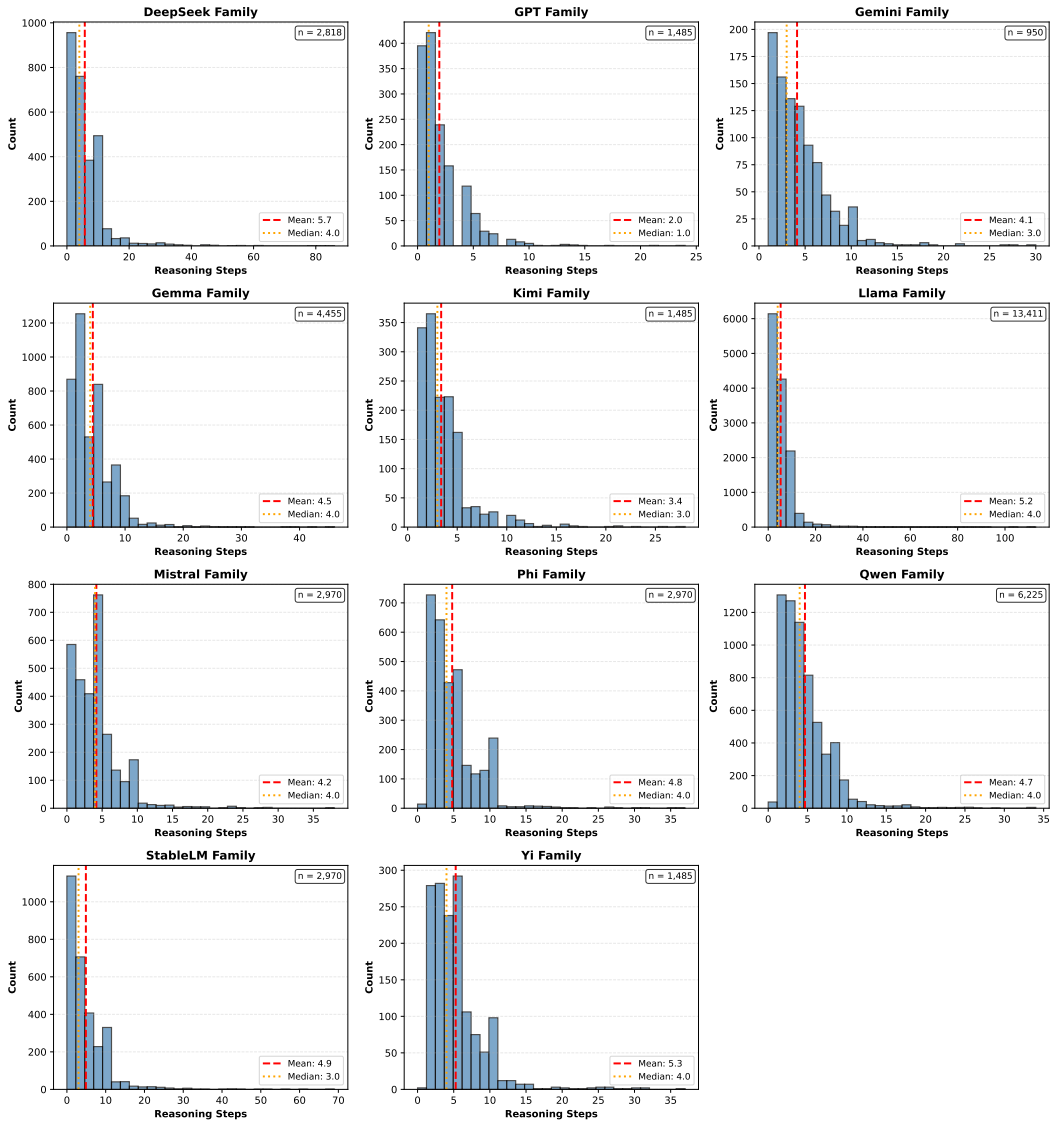



Figure 5: **Reasoning Length Distributions by Model Family.** Histograms show reasoning step counts for 11 model families across all benchmarks, with mean (red dashed) and median (orange dotted) lines. Most families cluster around 4–5 mean steps (Gemma: 4.5, Qwen: 4.7, Phi: 4.8, Mistral: 4.2). Notable exceptions: GPT shows most compact reasoning (mean 2.0, median 1.0), while Yi exhibits highest verbosity (mean 5.3). Llama displays widest distribution (range 0–100+ steps) due to spanning 1B–405B parameters. DeepSeek shows bimodal distribution (peaks at 4 and 20 steps, mean 5.7) reflecting architectural variation within the family.

### C.3 Per-Model Correlation Statistics.

Individual model analysis Table 2 reveals substantial heterogeneity in paradox manifestation across architectures and scales. Among 31 evaluated models, 27 (87%) exhibit negative overall correlations, confirming the paradox as the dominant pattern. The strongest negative GSM8K correlations appear in small-to-medium Gemma models (Gemma-2-2B: $\rho = -0.37$; Gemma-2-9B: $\rho = -0.31$), Microsoft Phi models (Phi-3-Mini: $\rho = -0.32$), and Mistral variants (Mistral-7B: $\rho = -0.29$). Notable exceptions include Llama-2-13B and GPT-OSS-20B, which show strong positive *overall* correlations ($\rho = +0.30$ and $+0.27$ respectively), driven by retrieval and commonsense tasks (Llama-2-13B: ARC-Easy $\rho = +0.54$, CommonsenseQA $+0.44$; GPT-OSS-20B: BoolQ $+0.80$, CommonsenseQA $+0.49$). DeepSeek models demonstrate architecture-dependent patterns: the 7B base model shows strong positive correlations on retrieval

Table 2: **Per-Model Correlation Statistics.** Individual model correlations between reasoning steps and accuracy. Missing values indicate insufficient samples (<10) for reliable estimation. Architecture: GQA (Grouped Query Attention), Std (standard multi-head attention). Sample sizes vary by evaluation coverage. $n = 31$ models.

| **Model** | **Params (B)** | **Arch** | **Family** | **GSM8K** $\rho$ | **BoolQ** $\rho$ | **ARC-E** $\rho$ | **CSQ** $\rho$ | **MMLU** $\rho$ | **Overall** $\rho$ | $n$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Qwen2.5-0.5B | 0.5 | Std | Qwen | $-0.20$ | $+0.05$ | $-0.03$ | $-0.06$ | $-0.03$ | $-0.13$ | 1485 |
| Gemma-3-1B | 1.0 | Std | Gemma | $-0.23$ | $+0.04$ | $-0.06$ | $-0.03$ | $+0.04$ | $-0.17$ | 1485 |
| Llama-3.2-1B | 1.0 | GQA | Llama | $-0.22$ | $-0.14$ | $-0.04$ | $-0.05$ | $+0.01$ | $-0.09$ | 1485 |
| StableLM-2-1.6B | 1.6 | Std | StableLM | $-0.13$ | $-0.22$ | $-0.01$ | $+0.09$ | $-0.06$ | $-0.27$ | 1485 |
| Gemma-2-2B | 2.0 | Std | Gemma | $-0.37$ | $+0.10$ | $-0.05$ | $+0.06$ | $-0.10$ | $-0.15$ | 1485 |
| Llama-3.2-3B | 3.0 | GQA | Llama | $-0.18$ | $-0.06$ | $-0.10$ | $-0.05$ | $-0.14$ | $-0.15$ | 1485 |
| StableLM-Zephyr-3B | 3.0 | Std | StableLM | $-0.20$ | $+0.01$ | $-0.01$ | $-0.03$ | $-0.05$ | $-0.19$ | 1485 |
| Qwen2.5-3B | 3.0 | Std | Qwen | $-0.20$ | $+0.00$ | $-0.08$ | $-0.10$ | $-0.02$ | $-0.08$ | 1485 |
| Phi-3-Mini | 3.8 | Std | Phi | $-0.32$ | $-0.01$ | $-0.06$ | $+0.10$ | $-0.01$ | $-0.07$ | 1485 |
| Phi-3.5-Mini | 3.8 | Std | Phi | $-0.29$ | $-0.05$ | $+0.05$ | $+0.06$ | $-0.06$ | $-0.09$ | 1485 |
| Minitron-4B-Width | 4.0 | GQA | Llama | $-0.25$ | $-0.00$ | $-0.13$ | $-0.17$ | $-0.14$ | $-0.23$ | 1485 |
| Minitron-4B-Depth | 4.0 | GQA | Llama | $-0.06$ | $-0.01$ | $-0.18$ | $-0.09$ | $-0.02$ | $-0.27$ | 1485 |
| Yi-1.5-6B | 6.0 | Std | Yi | $-0.05$ | $+0.14$ | $+0.02$ | $-0.01$ | $-0.02$ | $-0.01$ | 1485 |
| Mistral-7B-v0.3 | 7.0 | Std | Mistral | $-0.29$ | $-0.09$ | $-0.03$ | $-0.02$ | $-0.02$ | $-0.24$ | 1485 |
| DeepSeek-LLM-7B | 7.0 | Std | DeepSeek | $-0.01$ | $-0.05$ | $+0.54$ | $+0.18$ | $+0.24$ | $+0.16$ | 1485 |
| Qwen2.5-7B | 7.0 | Std | Qwen | $-0.24$ | $+0.02$ | $-0.03$ | $+0.01$ | $+0.00$ | $-0.10$ | 1485 |
| Llama-3.1-8B | 8.0 | GQA | Llama | $-0.09$ | $+0.03$ | $-0.02$ | $-0.08$ | $-0.08$ | $-0.11$ | 1485 |
| Nemotron-Nano-8B | 8.0 | GQA | Llama | $-0.08$ | $-0.07$ | $-0.14$ | $-0.05$ | $-0.07$ | $-0.06$ | 1485 |
| Gemma-2-9B | 9.0 | Std | Gemma | $-0.31$ | $+0.59$ | $+0.03$ | $+0.06$ | $+0.13$ | $-0.05$ | 1485 |
| Llama-2-13B | 13.0 | Std | Llama | $+0.05$ | $+0.36$ | $+0.54$ | $+0.44$ | — | $+0.30$ | 1200 |
| Qwen2.5-14B | 14.0 | Std | Qwen | $-0.16$ | $+0.02$ | $-0.05$ | $-0.11$ | $-0.04$ | $-0.05$ | 1770 |
| GPT-OSS-20B | 20.0 | Std | GPT | $+0.17$ | $+0.80$ | $+0.02$ | $+0.49$ | $+0.41$ | $+0.27$ | 1485 |
| Mistral-Small-24B | 24.0 | Std | Mistral | $-0.26$ | $+0.00$ | $-0.11$ | $+0.05$ | $-0.09$ | $-0.16$ | 1485 |
| Kimi-K2-32B | 32.0 | Std | Kimi | $-0.22$ | — | $+0.03$ | $-0.03$ | $+0.15$ | $-0.20$ | 1485 |
| Qwen2.5-32B | 32.0 | Std | Qwen | $-0.21$ | $+0.05$ | $-0.05$ | $+0.03$ | $-0.18$ | $-0.25$ | 1485 |
| Gemini-2.0-Flash | 50.0 | Std | Gemini | $-0.30$ | $-0.21$ | $-0.12$ | $-0.10$ | $-0.04$ | $-0.23$ | 950 |
| DeepSeek-LLM-67B | 67.0 | GQA | DeepSeek | $-0.18$ | $-0.02$ | $+0.04$ | $+0.26$ | $+0.05$ | $-0.02$ | 1333 |
| Llama-3-70B | 70.0 | GQA | Llama | $-0.37$ | $+0.00$ | $+0.73$ | $-0.19$ | $+0.44$ | $+0.01$ | 883 |
| Llama-3-70B-Inst | 70.0 | GQA | Llama | $-0.24$ | $-0.31$ | $+0.14$ | $-0.04$ | $-0.05$ | $-0.08$ | 933 |
| Llama-3.3-70B | 70.0 | GQA | Llama | $-0.12$ | — | $+0.05$ | $+0.02$ | $-0.03$ | $-0.35$ | 1397 |
| Llama-3.1-405B | 405.0 | GQA | Llama | $-0.16$ | $-0.25$ | $-0.04$ | $-0.03$ | $+0.01$ | $-0.31$ | 1485 |

tasks (ARC-Easy: $\rho = +0.54$), while the 67B model exhibits near-zero overall correlation ($\rho = -0.02$), suggesting that paradox strength varies with scale even within a single family. The largest evaluated model (Llama-3.1-405B, 405B parameters) maintains strong negative overall correlation ($\rho = -0.31$), indicating parameter scaling alone does not resolve the efficiency constraint. Qwen family models show consistent moderate negative correlations across all scales (0.5B to 32B: $\rho$ range $-0.05$ to $-0.25$), demonstrating family-level training methodology strongly influences paradox strength independent of parameter count. Missing values in BoolQ column for Kimi-K2-32B and Llama-3.3-70B, and in MMLU-STEM for Llama-2-13B reflect evaluation data availability constraints rather than fundamental model limitations.

### C.4 Question and Model Characteristic Correlations with Accuracy

Table 3: **Question and Model Characteristic Correlations with Accuracy.** Correlations between task/model characteristics and final answer accuracy across 44,106 evaluations. Question Length measured in words; Syntactic Complexity computed as weighted sum of clause count and conditional statement presence; Has Numbers indicates presence of numerical values in question text (analyzed via point-biserial correlation due to binary nature); Required Steps represents reasoning chain length extracted from model-generated traces; Trace Length measured in approximate tokens via word count aggregation. All correlations computed via Spearman rank correlation except Has Numbers (point-biserial). Trace Length shows the strongest negative correlation ($\rho = -0.146$), followed by Required Steps ($\rho = -0.093$, $p < 0.001$), confirming the step-accuracy paradox persists independent of question surface characteristics.

| **Characteristic** | **Correlation ($\rho$)** | **$p$-value** | **Test Type** | $n$ |
|---|---|---|---|---|
| Question Length (words) | $+0.010$ | 0.041 | Spearman | 44,106 |
| Syntactic Complexity | $+0.035$ | $< 0.001$ | Spearman | 44,106 |
| Has Numbers | $-0.076$ | $< 0.001$ | Point-Biserial | 44,106 |
| Required Steps | $-0.093$ | $< 0.001$ | Spearman | 44,106 |
| Trace Length (tokens) | $-0.146$ | $< 0.001$ | Spearman | 44,106 |

Question characteristic analysis in Table 3 establishes that the step-accuracy paradox represents a genuine reasoning efficiency constraint rather than confounding from question difficulty or surface features. Question Length and Syntactic Complexity show near-zero correlations with accuracy ($\rho = +0.010$ and $+0.035$ respectively), indicating accuracy is largely independent of question verbosity and structural complexity. The weak negative correlation for Has Numbers ($\rho = -0.076$) suggests numerical questions pose slightly greater difficulty, consistent with GSM8K's mathematical focus showing strongest paradox effects. Critically, Required Steps exhibits a clear negative correlation ($\rho = -0.093$), demonstrating the paradox magnitude matches the overall pattern reported in main text. Trace Length (tokens) shows even stronger negative correlation ($\rho = -0.146$, $1.6\times$ magnitude of step count), revealing that verbosity beyond structural decomposition contributes additional accuracy degradation. The $1.6\times$ amplification from steps to tokens suggests each reasoning step in longer chains contains increased within-step verbosity, compounding the efficiency penalty. Comparison between Required Steps and question characteristics confirms the paradox is not an artifact of task difficulty: if longer reasoning simply indicated harder problems (which happen to have lower accuracy), Question Length and Complexity would correlate strongly with both Steps and Accuracy, creating a confounding triangle. Instead, negligible question-accuracy correlations ($|\rho| < 0.03$) combined with moderate steps-accuracy correlation ($\rho = -0.093$) isolates the paradox as a reasoning process inefficiency independent of inherent problem difficulty. This pattern validates the interpretation that extended reasoning chains represent quality degradation rather than appropriate responses to genuinely harder questions.

### C.5 Correlation Matrix of Question and Model Characteristics

The correlation matrix in Figure 6 establishes three critical findings regarding paradox mechanisms and confound elimination. First, the strong correlation between Reasoning Steps and Trace Length ($\rho = 0.732$) with correlation less than unity confirms these represent related but distinct efficiency dimensions: step count captures reasoning structure while token length captures verbosity density. The differential accuracy correlations (Trace: $-0.146$ vs. Steps: $-0.093$, ratio $1.6\times$) reveal that

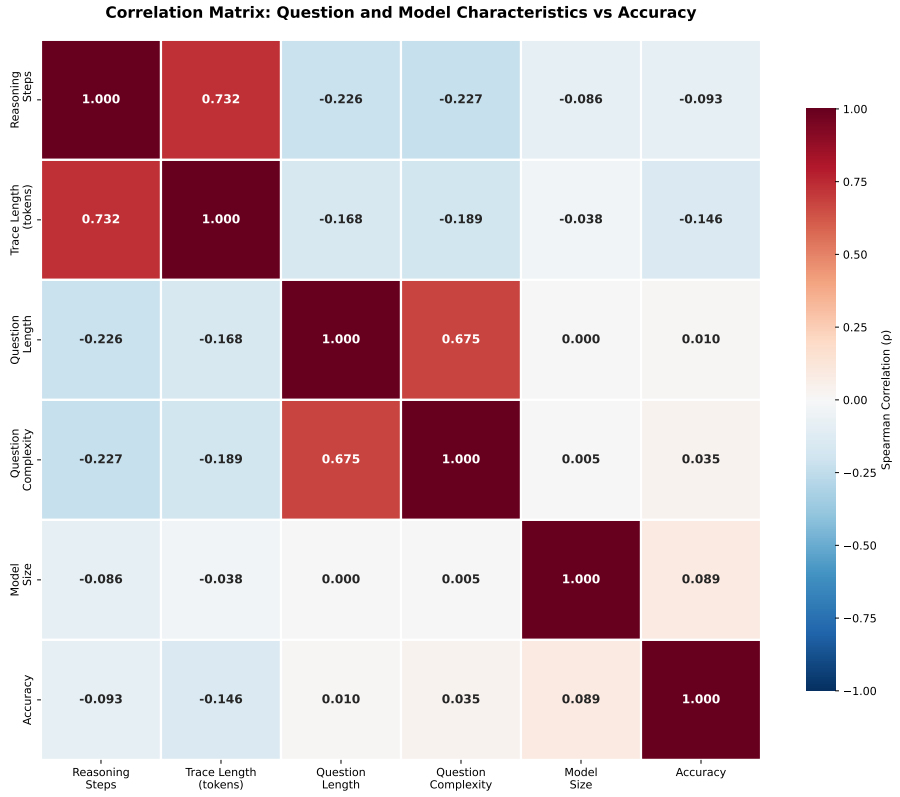



Figure 6: **Correlation Matrix of Question and Model Characteristics.** Spearman correlation heatmap displays pairwise relationships among six variables across 44,106 evaluations. Color intensity indicates correlation strength (dark red: strong positive; dark blue: strong negative; white: near-zero). Reasoning Steps and Trace Length (tokens) demonstrate high positive correlation ($\rho = 0.732$), confirming these metrics capture related but distinct aspects of verbosity: structural decomposition (steps) versus absolute output length (tokens). Both correlate negatively with Accuracy (Steps: $\rho = -0.093$; Trace Length: $\rho = -0.146$), validating the step-accuracy paradox through independent measurement approaches. Question Length and Question Complexity exhibit strong positive correlation ($\rho = 0.675$), establishing syntactic complexity increases with question verbosity as expected. Model Size shows weak positive correlation with Accuracy ($\rho = +0.089$), confirming parameter scaling provides modest performance improvements consistent with established scaling laws.

within-step verbosity contributes independent accuracy penalty beyond structural decomposition effects, suggesting both reasoning granularity and expression conciseness independently influence performance. Second, near-zero correlations between question characteristics (Length, Complexity) and Accuracy ($\rho = +0.010$ and $+0.035$ respectively) argue against question difficulty as a confounding variable in the step-accuracy paradox. If the paradox simply reflected harder questions requiring both longer reasoning and achieving lower accuracy, Question Complexity would correlate strongly with both Reasoning Steps and Accuracy, creating a confounding triangle. Instead, the matrix shows Question Complexity correlates weakly with Steps ($\rho = -0.227$) but negligibly with Accuracy ($\rho = +0.035$), breaking the confounding chain. This pattern isolates Reasoning Steps as an independent predictor of accuracy degradation unrelated to inherent problem difficulty. Third, the negative correlations between Reasoning/Trace Length and Question metrics ($\rho \approx -0.17$ to $-0.23$, all negative) reveal counterintuitive verbosity patterns: models generate longer reasoning chains for simpler, shorter questions rather than complex, lengthy ones. This inverse relationship supports interpretation of extended reasoning as compensatory behavior for capability limitations rather than appropriate responses to genuinely challenging problems. The weak Model Size–Accuracy correlation ($\rho = +0.089$) combined with stronger Reasoning Steps–Accuracy correlation ($\rho = -0.093$) demonstrates that reasoning efficiency (step economy) rivals parameter scaling as a performance determinant, with similar effect magnitudes ($|\rho| \approx 0.09$) despite representing fundamentally different model properties. Overall, the matrix validates the step-accuracy paradox as a robust efficiency

constraint independent of question difficulty, model scale, or measurement artifacts, establishing extended reasoning as a genuine failure mode rather than a statistical confound.

# D Near-Universal convergence at 1,000 tokens

## D.1 Complete Token Ablation Results Of Standard Models

Wilson score confidence intervals used throughout for improved coverage with extreme proportions. Sample sizes ($n = 50$ for most conditions) ensure adequate statistical power for plateau detection.

Table 4: **Complete token ablation results across all model-dataset combinations.** All 96 conditions (8 models $\times$ 3 datasets $\times$ 4 token limits) with accuracies, 95% Wilson confidence intervals, and sample sizes. Results demonstrate near-universal convergence at 1000–2000 token boundary across model scales (1B–405B parameters) and task types (mathematical reasoning, binary comprehension, science knowledge). Perfect plateaus ($\Delta = 0.0$pp, $p = 1.00$) occur in 21 of 24 model-dataset combinations (87.5%). The three exceptions—all involving Llama-3.1-405B—show small non-significant differences: GSM8K $+3.0$pp ($p = 0.506$), BoolQ $+0.0$pp ($p = 1.00$), ARC-Easy $+2.0$pp ($p = 0.558$).

| Model | Dataset | Token Limit | Accuracy (%) | 95% CI | n |
|---|---|---|---|---|---|
| Llama-3.2-1B | ARC-EASY | 300 | 72.0 | [58.3, 82.5] | 50 |
| Llama-3.2-1B | ARC-EASY | 500 | 72.0 | [58.3, 82.5] | 50 |
| Llama-3.2-1B | ARC-EASY | 1000 | 72.0 | [58.3, 82.5] | 50 |
| Llama-3.2-1B | ARC-EASY | 2000 | 72.0 | [58.3, 82.5] | 50 |
| Llama-3.2-1B | BOOLQ | 300 | 32.0 | [20.8, 45.8] | 50 |
| Llama-3.2-1B | BOOLQ | 500 | 32.0 | [20.8, 45.8] | 50 |
| Llama-3.2-1B | BOOLQ | 1000 | 32.0 | [20.8, 45.8] | 50 |
| Llama-3.2-1B | BOOLQ | 2000 | 32.0 | [20.8, 45.8] | 50 |
| Llama-3.2-1B | GSM8K | 300 | 26.0 | [15.9, 39.6] | 50 |
| Llama-3.2-1B | GSM8K | 500 | 26.0 | [15.9, 39.6] | 50 |
| Llama-3.2-1B | GSM8K | 1000 | 26.0 | [15.9, 39.6] | 50 |
| Llama-3.2-1B | GSM8K | 2000 | 26.0 | [15.9, 39.6] | 50 |
| Qwen2.5-3B | ARC-EASY | 300 | 84.0 | [71.5, 91.7] | 50 |
| Qwen2.5-3B | ARC-EASY | 500 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-3B | ARC-EASY | 1000 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-3B | ARC-EASY | 2000 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-3B | BOOLQ | 300 | 58.0 | [44.2, 70.6] | 50 |
| Qwen2.5-3B | BOOLQ | 500 | 58.0 | [44.2, 70.6] | 50 |
| Qwen2.5-3B | BOOLQ | 1000 | 58.0 | [44.2, 70.6] | 50 |
| Qwen2.5-3B | BOOLQ | 2000 | 58.0 | [44.2, 70.6] | 50 |
| Qwen2.5-3B | GSM8K | 300 | 58.0 | [44.2, 70.6] | 50 |
| Qwen2.5-3B | GSM8K | 500 | 68.0 | [54.2, 79.2] | 50 |
| Qwen2.5-3B | GSM8K | 1000 | 72.0 | [58.3, 82.5] | 50 |
| Qwen2.5-3B | GSM8K | 2000 | 72.0 | [58.3, 82.5] | 50 |
| Mistral-7B | ARC-EASY | 300 | 86.0 | [73.8, 93.0] | 50 |
| Mistral-7B | ARC-EASY | 500 | 86.0 | [73.8, 93.0] | 50 |
| Mistral-7B | ARC-EASY | 1000 | 86.0 | [73.8, 93.0] | 50 |
| Mistral-7B | ARC-EASY | 2000 | 86.0 | [73.8, 93.0] | 50 |
| Mistral-7B | BOOLQ | 300 | 84.0 | [71.5, 91.7] | 50 |
| Mistral-7B | BOOLQ | 500 | 84.0 | [71.5, 91.7] | 50 |
| Mistral-7B | BOOLQ | 1000 | 84.0 | [71.5, 91.7] | 50 |
| Mistral-7B | BOOLQ | 2000 | 84.0 | [71.5, 91.7] | 50 |
| Mistral-7B | GSM8K | 300 | 30.0 | [19.1, 43.8] | 50 |
| Mistral-7B | GSM8K | 500 | 32.0 | [20.8, 45.8] | 50 |
| Mistral-7B | GSM8K | 1000 | 32.0 | [20.8, 45.8] | 50 |
| Mistral-7B | GSM8K | 2000 | 32.0 | [20.8, 45.8] | 50 |

Continued on next page

Table 4 – continued from previous page

| Model | Dataset | Token Limit | Accuracy (%) | 95% CI | n |
|---|---|---|---|---|---|
| Qwen2.5-7B | ARC-EASY | 300 | 90.0 | [78.6, 95.7] | 50 |
| Qwen2.5-7B | ARC-EASY | 500 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-7B | ARC-EASY | 1000 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-7B | ARC-EASY | 2000 | 92.0 | [81.2, 96.8] | 50 |
| Qwen2.5-7B | BOOLQ | 300 | 78.0 | [64.8, 87.2] | 50 |
| Qwen2.5-7B | BOOLQ | 500 | 78.0 | [64.8, 87.2] | 50 |
| Qwen2.5-7B | BOOLQ | 1000 | 78.0 | [64.8, 87.2] | 50 |
| Qwen2.5-7B | BOOLQ | 2000 | 78.0 | [64.8, 87.2] | 50 |
| Qwen2.5-7B | GSM8K | 300 | 54.0 | [40.4, 67.0] | 50 |
| Qwen2.5-7B | GSM8K | 500 | 78.0 | [64.8, 87.2] | 50 |
| Qwen2.5-7B | GSM8K | 1000 | 82.0 | [69.2, 90.2] | 50 |
| Qwen2.5-7B | GSM8K | 2000 | 82.0 | [69.2, 90.2] | 50 |
| Llama-3.1-8B | ARC-EASY | 300 | 92.0 | [81.2, 96.8] | 50 |
| Llama-3.1-8B | ARC-EASY | 500 | 94.0 | [83.8, 97.9] | 50 |
| Llama-3.1-8B | ARC-EASY | 1000 | 94.0 | [83.8, 97.9] | 50 |
| Llama-3.1-8B | ARC-EASY | 2000 | 94.0 | [83.8, 97.9] | 50 |
| Llama-3.1-8B | BOOLQ | 300 | 84.0 | [71.5, 91.7] | 50 |
| Llama-3.1-8B | BOOLQ | 500 | 84.0 | [71.5, 91.7] | 50 |
| Llama-3.1-8B | BOOLQ | 1000 | 84.0 | [71.5, 91.7] | 50 |
| Llama-3.1-8B | BOOLQ | 2000 | 84.0 | [71.5, 91.7] | 50 |
| Llama-3.1-8B | GSM8K | 300 | 54.0 | [40.4, 67.0] | 50 |
| Llama-3.1-8B | GSM8K | 500 | 48.0 | [34.8, 61.5] | 50 |
| Llama-3.1-8B | GSM8K | 1000 | 44.0 | [31.2, 57.7] | 50 |
| Llama-3.1-8B | GSM8K | 2000 | 44.0 | [31.2, 57.7] | 50 |
| Gemma-2-9B | ARC-EASY | 300 | 94.0 | [83.8, 97.9] | 50 |
| Gemma-2-9B | ARC-EASY | 500 | 94.0 | [83.8, 97.9] | 50 |
| Gemma-2-9B | ARC-EASY | 1000 | 94.0 | [83.8, 97.9] | 50 |
| Gemma-2-9B | ARC-EASY | 2000 | 94.0 | [83.8, 97.9] | 50 |
| Gemma-2-9B | BOOLQ | 300 | 82.0 | [69.2, 90.2] | 50 |
| Gemma-2-9B | BOOLQ | 500 | 82.0 | [69.2, 90.2] | 50 |
| Gemma-2-9B | BOOLQ | 1000 | 82.0 | [69.2, 90.2] | 50 |
| Gemma-2-9B | BOOLQ | 2000 | 82.0 | [69.2, 90.2] | 50 |
| Gemma-2-9B | GSM8K | 300 | 70.0 | [56.2, 80.9] | 50 |
| Gemma-2-9B | GSM8K | 500 | 70.0 | [56.2, 80.9] | 50 |
| Gemma-2-9B | GSM8K | 1000 | 70.0 | [56.2, 80.9] | 50 |
| Gemma-2-9B | GSM8K | 2000 | 70.0 | [56.2, 80.9] | 50 |
| Qwen2.5-72B | ARC-EASY | 300 | 84.0 | [71.5, 91.7] | 50 |
| Qwen2.5-72B | ARC-EASY | 500 | 98.0 | [89.5, 99.6] | 50 |
| Qwen2.5-72B | ARC-EASY | 1000 | 100.0 | [92.9, 100.0] | 50 |
| Qwen2.5-72B | ARC-EASY | 2000 | 100.0 | [92.9, 100.0] | 50 |
| Qwen2.5-72B | BOOLQ | 300 | 86.0 | [73.8, 93.0] | 50 |
| Qwen2.5-72B | BOOLQ | 500 | 86.0 | [73.8, 93.0] | 50 |
| Qwen2.5-72B | BOOLQ | 1000 | 86.0 | [73.8, 93.0] | 50 |
| Qwen2.5-72B | BOOLQ | 2000 | 86.0 | [73.8, 93.0] | 50 |
| Qwen2.5-72B | GSM8K | 300 | 56.0 | [42.3, 68.8] | 50 |
| Qwen2.5-72B | GSM8K | 500 | 88.0 | [76.2, 94.4] | 50 |
| Qwen2.5-72B | GSM8K | 1000 | 96.0 | [86.5, 98.9] | 50 |
| Qwen2.5-72B | GSM8K | 2000 | 96.0 | [86.5, 98.9] | 50 |
| Llama-3.1-405B | ARC-EASY | 300 | 35.0 | [26.4, 44.7] | 100 |
| Llama-3.1-405B | ARC-EASY | 500 | 100.0 | [92.9, 100.0] | 50 |
| Llama-3.1-405B | ARC-EASY | 1000 | 96.0 | [86.5, 98.9] | 50 |
| Llama-3.1-405B | ARC-EASY | 2000 | 98.0 | [89.5, 99.6] | 50 |
| Llama-3.1-405B | BOOLQ | 300 | 88.0 | [80.2, 93.0] | 100 |

Continued on next page

Table 4 – continued from previous page

| Model | Dataset | Token Limit | Accuracy (%) | 95% CI | n |
|---|---|---|---|---|---|
| Llama-3.1-405B | BOOLQ | 500 | 91.0 | [83.8, 95.2] | 100 |
| Llama-3.1-405B | BOOLQ | 1000 | 87.0 | [79.0, 92.2] | 100 |
| Llama-3.1-405B | BOOLQ | 2000 | 87.0 | [79.0, 92.2] | 100 |
| Llama-3.1-405B | GSM8K | 300 | 76.0 | [66.8, 83.3] | 100 |
| Llama-3.1-405B | GSM8K | 500 | 82.0 | [73.3, 88.3] | 100 |
| Llama-3.1-405B | GSM8K | 1000 | 87.0 | [79.0, 92.2] | 100 |
| Llama-3.1-405B | GSM8K | 2000 | 90.0 | [82.6, 94.5] | 100 |

### D.1.1 Plateau Summary

Of the 21 plateau conditions, 8 represent genuine convergence: accuracy improves between 300 and 1,000 tokens before stabilising, indicating that the model benefits from additional tokens up to the boundary but not beyond. Twelve show flat trajectories from the lowest token limit, consistent with capacity limitation independent of token budget, and one (Llama-3.1-8B on GSM8K) declines before stabilising at 1,000 tokens. None of the 21 shows any gain beyond 1,000 tokens. The three exceptions—all involving Llama-3.1-405B Grattafiori et al. (2024)—show small, non-significant differences ($\Delta \leq +3.0$ pp, $P \geq 0.506$) and are visualised in Figure 7 and Table 4.

Figure 7 visualises the plateau classification for all 24 model–dataset conditions. Green rows indicate perfect plateau ($\Delta = 0.0$ pp, $P = 1.00$); yellow rows indicate small non-significant differences. All three Llama-3.1-405B conditions are the sole exceptions, consistent with the main text analysis.

**Plateau Summary**

**A. Plateau Detection Summary: 24 Model-Dataset Combinations**

| Model | Dataset | Accuracy @1000 tokens | Accuracy @2000 tokens | Change (pp) | p-value |
|---|---|---|---|---|---|
| Llama-3.2-1B | GSM8K | 26.0 | 26.0 | +0.0 | 1.00 |
| Llama-3.2-1B | BoolQ | 32.0 | 32.0 | +0.0 | 1.00 |
| Llama-3.2-1B | ARC-Easy | 72.0 | 72.0 | +0.0 | 1.00 |
| Qwen2.5-3B | GSM8K | 72.0 | 72.0 | +0.0 | 1.00 |
| Qwen2.5-3B | BoolQ | 58.0 | 58.0 | +0.0 | 1.00 |
| Qwen2.5-3B | ARC-Easy | 92.0 | 92.0 | +0.0 | 1.00 |
| Mistral-7B | GSM8K | 32.0 | 32.0 | +0.0 | 1.00 |
| Mistral-7B | BoolQ | 84.0 | 84.0 | +0.0 | 1.00 |
| Mistral-7B | ARC-Easy | 86.0 | 86.0 | +0.0 | 1.00 |
| Qwen2.5-7B | GSM8K | 82.0 | 82.0 | +0.0 | 1.00 |
| Qwen2.5-7B | BoolQ | 78.0 | 78.0 | +0.0 | 1.00 |
| Qwen2.5-7B | ARC-Easy | 92.0 | 92.0 | +0.0 | 1.00 |
| Llama-3.1-8B | GSM8K | 44.0 | 44.0 | +0.0 | 1.00 |
| Llama-3.1-8B | BoolQ | 84.0 | 84.0 | +0.0 | 1.00 |
| Llama-3.1-8B | ARC-Easy | 94.0 | 94.0 | +0.0 | 1.00 |
| Gemma-2-9B | GSM8K | 70.0 | 70.0 | +0.0 | 1.00 |
| Gemma-2-9B | BoolQ | 82.0 | 82.0 | +0.0 | 1.00 |
| Gemma-2-9B | ARC-Easy | 94.0 | 94.0 | +0.0 | 1.00 |
| Qwen2.5-72B | GSM8K | 96.0 | 96.0 | +0.0 | 1.00 |
| Qwen2.5-72B | BoolQ | 86.0 | 86.0 | +0.0 | 1.00 |
| Qwen2.5-72B | ARC-Easy | 100.0 | 100.0 | +0.0 | 1.00 |
| Llama-3.1-405B | GSM8K | 87.0 | 90.0 | +3.0 | 0.506 |
| Llama-3.1-405B | BoolQ | 87.0 | 87.0 | +0.0 | 1.00 |
| Llama-3.1-405B | ARC-Easy | 96.0 | 98.0 | +2.0 | 0.558 |

Perfect Plateau (Δ=0.0pp, p=1.00) Small Non-Significant (p>0.05)

Figure 7: **Plateau detection summary across all 24 model–task conditions.** Green shading indicates perfect plateau ($\Delta = 0.0$ pp, $P = 1.00$; 21 of 24 conditions). Yellow shading indicates small non-significant differences (Llama-3.1-405B conditions only; $\Delta \leq +3.0$ pp, $P \geq 0.506$). Full statistics in Table 4 above.

## D.2 Token Ablation: BoolQ and ARC-Easy

Token ablation on BoolQ and ARC-Easy replicates the near-universal convergence boundary observed on GSM8K (main text Figure 1A). Across all eight models (1B–405B parameters), every model–task combination reaches stable accuracy at approximately 1,000 tokens (red dashed line), with no significant improvement beyond this boundary. BoolQ trajectories are flat from the lowest token limit

for most models, consistent with the binary response format requiring minimal elaboration. ARC-Easy shows steeper early improvement between 300 and 500 tokens—notably LLAMA-3.1-405B (35% to 100%)—but convergence is complete by 1,000 tokens across all models. Together with the GSM8K result, these patterns confirm that the 1,000-token boundary holds across qualitatively distinct task types, consistent with a general property of autoregressive reasoning.

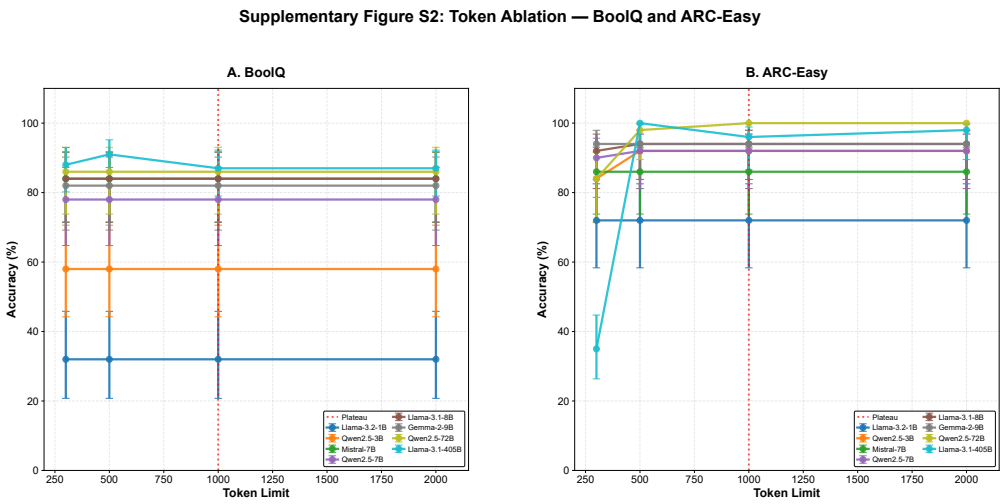



Figure 8: **Token ablation on BoolQ and ARC-Easy confirm near-universal convergence. (A)** BoolQ accuracy across eight standard models (1B–405B parameters) at token limits 300–2,000. All models plateau at or before 1,000 tokens (red dashed line); most trajectories are flat from 300 tokens, consistent with binary response format requiring little elaboration. **(B)** ARC-Easy accuracy across the same eight models. LLAMA-3.1-405B shows the steepest early improvement (35% at 300 tokens to 100% at 500 tokens) but converges by 1,000 tokens consistent with all other models. Error bars, 95% Wilson confidence intervals. The convergence boundary holds across three qualitatively distinct task types: mathematical (GSM8K; main text), binary linguistic (BoolQ), and knowledge retrieval (ARC-Easy).

### D.3 Complete causal intervention results

Models were used for the causal intervention are: Qwen2.5-32B-Instruct, Qwen2.5-3B-Instruct, Gemma-2B-IT, Llama-3.1-405B-Instruct, Llama-3.3-70B-Instruct, llama-3.2-3b-instruct and Qwen3-80B-A3B-Instruct. Negative gaps (Brief condition on GSM8K and MMLU-STEM) indicate *reversed* inverse scaling where large models outperform small models under constraints, demonstrating that over-generation disproportionately affects capable models. BoolQ shows minimal change ($-0.7$pp) likely due to yes/no format requiring minimal reasoning elaboration. Mean gap reduction across 5 datasets: 30.1pp (range: $-0.7$ to $+52.7$pp).

Table 5: **Complete causal intervention results across five datasets.** (A) Accuracy by model size and condition for all datasets tested. Small models include 2B–3B parameters; large models include 32B–405B parameters. (B) Gap reduction analysis showing how brief constraints reduce inverse scaling (small minus large model accuracy gap) across datasets. Overall gap reduced from $+38.3$pp to $+8.2$pp (78.6% relative reduction, 30.1pp absolute reduction), demonstrating robust generalization of the over-generation mechanism beyond GSM8K.

**(A) Accuracy Results by Dataset and Condition**

| Dataset | Condition | Small Acc (%) | Large Acc (%) | Gap (pp) |
|---|---|---|---|---|
| GSM8K | Control | 93.8 | 80.8 | +13.1 |
| GSM8K | Brief | 88.5 | 96.2 | -7.7 |
| BoolQ | Control | 77.9 | 54.4 | +23.5 |
| BoolQ | Brief | 82.4 | 58.1 | +24.3 |
| ARC-Easy | Control | 100.0 | 28.6 | +71.4 |
| ARC-Easy | Brief | 100.0 | 81.2 | +18.8 |
| CommonsenseQA | Control | 79.3 | 23.3 | +56.0 |
| CommonsenseQA | Brief | 67.2 | 45.7 | +21.6 |
| MMLU-STEM | Control | 50.0 | 22.7 | +27.3 |
| MMLU-STEM | Brief | 59.1 | 75.0 | -15.9 |

**(B) Gap Reduction Analysis**

| Dataset | Control Gap (pp) | Brief Gap (pp) | Gap Reduction (pp) |
|---|---|---|---|
| GSM8K | +13.1 | -7.7 | +20.8 |
| BoolQ | +23.5 | +24.3 | -0.7 |
| ARC-Easy | +71.4 | +18.8 | +52.7 |
| CommonsenseQA | +56.0 | +21.6 | +34.5 |
| MMLU-STEM | +27.3 | -15.9 | +43.2 |
| **Mean Reduction** | — | — | **+30.1** |

### D.4 Statistical validation of causal intervention effects

Table 6 presents paired $t$-tests comparing Brief versus Control conditions for large models on each dataset individually and pooled across all datasets. Each test pairs responses by both problem identity and model identity, ensuring that the same model answered the same problem under both conditions. The pooled test groups responses by unique problem identity (averaging across models within each dataset) to yield 115 independent matched observations. Four of five datasets show significant improvement under brief constraints ($P < 0.05$); BoolQ is the sole exception, consistent with its binary response format requiring minimal elaboration.

Table 6: **Statistical validation of brevity constraint effects across datasets.** Paired $t$-tests comparing Brief versus Control conditions for large models (32B–405B parameters). **Per-dataset tests** pair each response by problem and model identity ($n$ = number of matched response pairs). **Pooled test** averages across models within each problem, yielding 115 unique matched problems across all five datasets. Gap reduction = Control gap $-$ Brief gap, where gap = small-model accuracy $-$ large-model accuracy. Positive gap reduction indicates that brief constraints reduced inverse scaling. Significance threshold: $\alpha = 0.05$.

| Dataset | $n$ (pairs) | $t$ | df | $P$ | Gap reduction (pp) | Sig. |
|---|---|---|---|---|---|---|
| GSM8K | 52 | 2.68 | 51 | $1.0 \times 10^{-2}$ | $+20.8$ | Yes |
| ARC-Easy | 112 | 8.86 | 111 | $1.5 \times 10^{-14}$ | $+52.7$ | Yes |
| CommonsenseQA | 116 | 4.19 | 115 | $5.5 \times 10^{-5}$ | $+34.5$ | Yes |
| MMLU-STEM | 44 | 6.86 | 43 | $2.0 \times 10^{-8}$ | $+43.2$ | Yes |
| BoolQ | 136 | 1.00 | 135 | $3.2 \times 10^{-1}$ | $-0.7$ | No |
| **Pooled (all datasets)** | **115** | **7.59** | **114** | $\mathbf{9.3 \times 10^{-12}}$ | $\mathbf{+30.1}$ | **Yes** |

### D.5 Intervention prompts used in causal experiments

The Direct condition (not shown) requested answers without any reasoning, but exhibited anomalous performance on GSM8K (23% accuracy, likely due to prompt misinterpretation) and was excluded from analysis. Control and Brief conditions provide cleaner test of the over-generation hypothesis.

**MATH (GSM8K) — CONTROL**

```
Problem: {question}
Solution:
```

**MATH (GSM8K) — BRIEF (≤50 words)**

```
Problem: {question}
Provide a BRIEF solution in under 50 words. Show only the essential
calculation steps.
Solution:
```

**REASONING (BoolQ) — CONTROL**

```
Read the passage carefully and answer the question.
Passage: {passage}
Question: {question}
Think carefully about what the passage says. Answer with only "Yes" or "No".
Answer:
```

**REASONING (BoolQ) — BRIEF (≤10 words)**

```
Read the passage and answer.
Passage: {passage}
Question: {question}
Answer in 10 words or less: Yes or No, and why.
Answer:
```

**COMMONSENSE (ARC-Easy, CommonsenseQA) — CONTROL**

```
Question: {question}
Answer:
```

**COMMONSENSE — BRIEF (1 sentence)**

```
Answer this multiple choice question.
{question}
Answer with just the letter and ONE sentence explanation.
Answer:
```

### D.6 Random-sample validation and prompt-confound separation

The causal intervention in Section 3.1 used inverse-scaling problems—cases where large models underperform small models—raising two concerns: (1) regression to the mean may explain re-test gains on an adversarially selected sample, and (2) the brevity instruction constitutes a prompt change that could activate a different reasoning mode independently of token budget. We address both through a complementary experiment on randomly sampled problems with a three-condition design.

**Design.** We evaluated four models spanning three parameter scales— QWEN3-14B, QWEN3-32B, LLAMA-3.3-70B-INSTRUCT, and LLAMA-3.1-405B—on $n = 100$–$200$ problems drawn uniformly at random (seed=42, no pre-selection) from GSM8K and ARC-Easy under three conditions:

- **Control**: original prompt, unconstrained generation.
- **Brief**: brevity instruction in prompt (GSM8K: "Provide a BRIEF solution in under 50 words"; ARC-Easy: "Answer with just the letter and ONE sentence explanation"; Appendix D.5).
- **Length-only**: *identical* prompt to control; generation capped at the parameter level only (`max_new_tokens` $= 75$ for math, 30 for multiple-choice). No brevity instruction in prompt.

The length-only condition is the critical addition: because the prompt is byte-for-byte identical to control, any accuracy difference isolates the token budget effect from prompt framing. The *token budget effect* = Length-only − Control; the *framing effect* = Brief − Length-only.

**Results.** Table 7 reports per-model accuracy under all three conditions. Pairwise McNemar statistics appear in Table 8.

Table 7: **Random-sample causal intervention results.** Three conditions across four models and two datasets ($n = 100$–$200$ random problems per condition, seed=42). Token budget effect = length-only $\Delta$ vs control. Framing effect = brief − length-only. Pooled rows average across all models per dataset.

| Model | Dataset | $n$ | Control (%) | Brief (%) | Length-only (%) | Brief $\Delta$ | Framing effect |
|---|---|---|---|---|---|---|---|
| Llama-3.3-70B | ARC-Easy | 100 | 90.0 | 98.0 | 88.0 | +8.0pp | +10.0pp |
| Llama-3.3-70B | GSM8K | 100 | 85.0 | 91.0 | 86.0 | +6.0pp | +5.0pp |
| Llama-3.1-405B | ARC-Easy | 200 | 90.0 | 93.5 | 91.0 | +3.5pp | +2.5pp |
| Llama-3.1-405B | GSM8K | 200 | 88.0 | 91.5 | 89.5 | +3.5pp | +2.0pp |
| Qwen3-14B | ARC-Easy | 100 | 75.0 | 91.0 | 82.0 | +16.0pp | +9.0pp |
| Qwen3-14B | GSM8K | 100 | 83.0 | 86.0 | 82.0 | +3.0pp | +4.0pp |
| Qwen3-32B | ARC-Easy | 100 | 86.0 | 94.0 | 89.0 | +8.0pp | +5.0pp |
| Qwen3-32B | GSM8K | 100 | 79.0 | 90.0 | 82.0 | +11.0pp | +8.0pp |
| *Pooled* | ARC-Easy | — | 86.2 | 94.0 | 88.2 | +7.8pp | +5.8pp |
| *Pooled* | GSM8K | — | 84.6 | 90.0 | 85.8 | +5.4pp | +4.2pp |

**Claim 1 (constraining generation improves accuracy).** Brief constraints improve over control in all eight model–dataset conditions (+3.0–+16.0 pp), and length-only constraints improve over control in six of eight conditions (+1.0–+7.0 pp). The two exceptions (Llama-3.3-70B on ARC-Easy, −2 pp, $P = 0.79$; Qwen3-14B on GSM8K, −1 pp, $P = 1.00$) are not significant (Table 8). Pooled across all models, brief constraints yield +7.8 pp on ARC-Easy and +5.4 pp on GSM8K relative to unconstrained generation.

**Claim 2 (over-generation, not capacity exhaustion).** In six of eight conditions, imposing a token budget with an *unchanged* prompt improves accuracy over unconstrained generation (none individually significant), arguing against the alternative that models lack capacity at shorter budgets. The pattern is most pronounced at large scale: LLAMA-3.1-405B shows a framing effect of only 2.0–2.5 pp across both datasets, indicating that at 405B parameters both effects converge, with the token budget contributing +1.0–+1.5 pp and framing contributing +2.0–+2.5 pp.

**Mechanism separation.** The experiment reveals two additive contributions. Token budget alone contributes +1.0–+7.0 pp across conditions (six of eight positive); prompt framing contributes a further +2.0–+10.0 pp. Both mechanisms are positive in the majority of conditions, consistent with the over-generation interpretation: the token budget effect demonstrates that models possess the answer within a shorter budget, while the framing effect demonstrates that a conciseness signal prevents quality degradation under extended generation.

**Regression-to-mean exclusion.** Problems were drawn uniformly at random with no pre-selection based on Phase 1 performance, across two datasets and four models spanning 14B to 405B parameters. The positive effect of brevity constraints on a fresh random sample rules out regression to the mean as an alternative explanation.

**Summary.** Three alternative explanations for the main causal finding are addressed: (1) regression to the mean is excluded by the random-sample design across four models and two datasets; (2) prompt

Table 8: **Pairwise McNemar tests for random-sample causal intervention.** Paired by problem identity ($n$ matched pairs per row, two-tailed, $\alpha = 0.05$). *ns* = not significant. Framing tests marked **bold** where significant, (1 of 8), indicating that the brevity instruction can contribute independently of token budget.

| Model | Dataset | Comparison | $P$ | Sig. |
|---|---|---|---|---|
| Llama-3.3-70B | ARC-Easy | Control → Brief (+8pp) | 0.008 | * |
| | | Control → Length-only (−2pp) | 0.791 | ns |
| | | **Length-only vs Brief (+10pp)** | **0.006** | * |
| Llama-3.3-70B | GSM8K | Control → Brief (+6pp) | 0.210 | ns |
| | | Control → Length-only (+1pp) | 1.000 | ns |
| | | Length-only vs Brief (+5pp) | 0.332 | ns |
| Llama-3.1-405B | ARC-Easy | Control → Brief (+3.5pp) | 0.093 | ns |
| | | Control → Length-only (+1pp) | 0.694 | ns |
| | | Length-only vs Brief (+2.5pp) | 0.237 | ns |
| Llama-3.1-405B | GSM8K | Control → Brief (+3.5pp) | 0.054 | ns |
| | | Control → Length-only (+1.5pp) | 0.512 | ns |
| | | Length-only vs Brief (+2.0pp) | 0.302 | ns |
| Qwen3-14B | ARC-Easy | Control → Brief (+16pp) | 0.005 | * |
| | | Control → Length-only (+7pp) | 0.144 | ns |
| | | Length-only vs Brief (+9pp) | 0.108 | ns |
| Qwen3-14B | GSM8K | Control → Brief (+3pp) | 0.678 | ns |
| | | Control → Length-only (−1pp) | 1.000 | ns |
| | | Length-only vs Brief (+4pp) | 0.503 | ns |
| Qwen3-32B | ARC-Easy | Control → Brief (+8pp) | 0.115 | ns |
| | | Control → Length-only (+3pp) | 0.607 | ns |
| | | Length-only vs Brief (+5pp) | 0.332 | ns |
| Qwen3-32B | GSM8K | Control → Brief (+11pp) | 0.061 | ns |
| | | Control → Length-only (+3pp) | 0.648 | ns |
| | | Length-only vs Brief (+8pp) | 0.169 | ns |

framing and token budget are separated and both positive across the majority of conditions; (3) the effect generalises beyond the inverse-scaling subpopulation to randomly drawn problems, with brief constraints improving accuracy in all eight model–dataset conditions (+3.0–+16.0 pp over control; exact binomial sign test, one-sided: $P = 0.004$ (8/8 positive: $P = (1/2)^8 = 0.0039$); length-only: 6/8 positive, not significant, $P = 0.145$). At the largest scale tested (405B), the framing effect contracts to $\leq 2.5$ pp, consistent with the interpretation that highly capable models require only a token budget signal to suppress over-generation.

### D.7 Manual error categorization by intervention condition

Error categorization was performed independently by two annotators using a blinded protocol (condition labels removed; see annotation protocol below in D.10.4). Inter-annotator agreement was computed on 57 of 58 responses (one excluded from the agreement calculation due to ambiguous text extraction but included in the final categorization after joint review): Cohen's $\kappa = 0.886$ for error category (94.7% raw agreement) and $\kappa = 0.808$ for error location (89.5% raw agreement). Seven disagreements were resolved by joint review of the complete reasoning trace.

Errors were classified into two categories: **Premature Termination** (Incomplete) and **Error Propagation** (Logical). Under unconstrained generation, 76% of failures involved premature termination (→45% under constraints, −31 pp); error propagation rose symmetrically from 24% to 55% (+31 pp). This symmetric tradeoff—where constraints redirect failure mode from abandonment to forced conclusion—yields the net +15.4 pp accuracy improvement.

### D.8 Token Utilization Pattern of Standard Models

Figure 9 Token counts represent output tokens only (excluding input prompt). Control condition shows high variability (CV ≈ 0.7) indicating inconsistent verbosity across problems, while Brief shows low variability (CV ≈ 0.4) confirming constraint adherence. The lack of strong positive

Table 9: **Manual error categorization by intervention condition.** Distribution of error types across Control (verbose, $n = 29$) and Brief (constrained, $n = 29$) conditions on GSM8K. Errors were categorized by type based on systematic review of model reasoning traces. **Incomplete** (premature termination): Models abandoned reasoning before reaching solution despite available tokens. **Logical**: Error propagation through faulty reasoning steps.

| Error Category | Control (n) | Control (%) | Brief (n) | Brief (%) |
|---|---|---|---|---|
| Incomplete (Premature) | 22 | 76 | 13 | 45 |
| Logical (Error Prop.) | 7 | 24 | 16 | 55 |
| **Total** | **29** | **100.0** | **29** | **100.0** |

correlation between length and accuracy in Control is critical evidence against the hypothesis that models benefit from additional reasoning tokens—instead, very long responses (>400 tokens) show degraded accuracy, suggesting meandering behavior. Response length validation: Control mean = 202 words, Brief mean = 81 words (59.8% reduction as reported in main text Section 2.1).

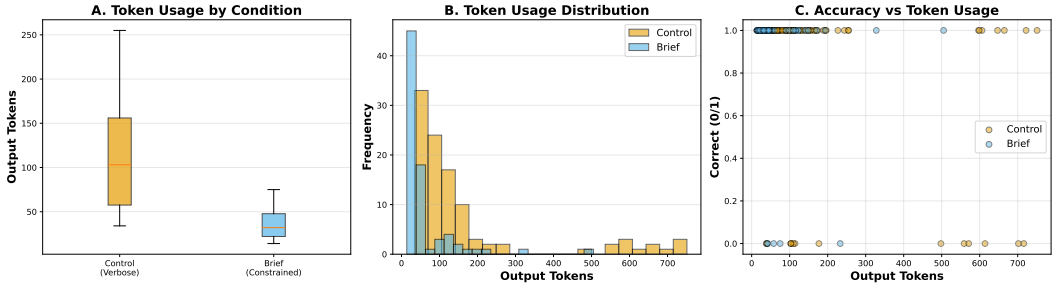



Figure 9: **Token utilization patterns validate manipulation success and reveal early termination behavior.** Data from GSM8K causal intervention experiment ($n = 182$ responses across 7 models). **(A)** Boxplots show Control condition generates substantially longer responses than Brief (median ≈155 vs ≈48 tokens, IQR 59–256 vs 25–78). Brief constraints successfully reduced output token count by ≈69% without complete suppression of reasoning. Whiskers extend to 1.5×IQR; outliers not shown. **(B)** Distribution histograms confirm successful constraint adherence: Brief responses tightly clustered around 50-token target (mode ≈45 tokens), while Control exhibits wide variation (range 35–750 tokens, mode ≈90 tokens). Bimodal distribution in Control suggests two response strategies: concise (<100 tokens) and verbose (>150 tokens). **(C)** Scatter plot reveals no clear positive relationship between response length and accuracy in Control condition (visual inspection suggests weak positive trend for brief responses but degradation beyond 200 tokens). Brief constraints improve accuracy despite reduced length (orange vs blue clusters at $y = 1$), contradicting capacity exhaustion hypothesis and supporting over-generation mechanism.

### D.9 Model-specific error patterns

Figure 10 displays per-model error distributions under Control (Panel A) and Brief (Panel B) conditions using the two-category annotation scheme (incomplete vs logical). Seven models contributed errors across both conditions: Qwen2.5-32B-Instruct, Qwen2.5-3B-Instruct, Gemma-2B-IT, Llama-3.1-405B-Instruct, Llama-3.3-70B-Instruct, llama-3.2-3b-instruct and Qwen3-80B-Instruct. **Llama-3.2-3B-Instruct** errors appear exclusively under the Brief condition (Panel B); it contributed no errors under Control, and therefore does not appear in Panel A. Under Control, most models show incomplete-dominant error profiles (cyan bars >60%), consistent with the aggregate 76% premature termination rate. Under Brief constraints, the incomplete proportion shrinks universally while logical errors (yellow) expand, confirming that the tradeoff generalises across architectures (Llama, Qwen, Gemma) and scales (2B–405B).

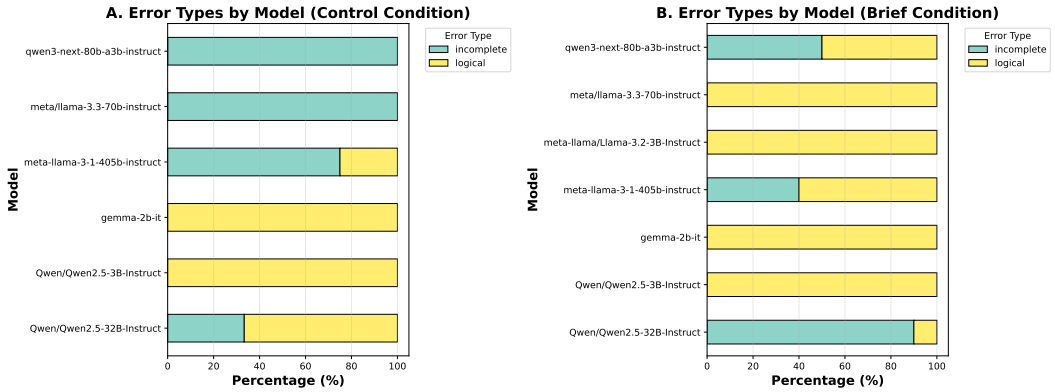



Figure 10: **Model-specific error patterns under two-category blinded annotation.** Error distribution across seven models under (A) Control and (B) Brief conditions. **Incomplete** (cyan): premature termination despite available tokens. **Logical** (yellow): error propagation through faulty reasoning. All models show reduction in premature termination under Brief constraints (cyan regions shrink from Panel A to B), with the aggregate shift from 76% to 45% incomplete errors ($-31$ pp). Inter-annotator agreement: Cohen's $\kappa = 0.886$ with condition labels hidden from the second annotator.

### D.10 Statistical Methods and Additional Notes

#### D.10.1 Token Ablation Analysis (Section D.1, Table 4)

**Plateau Detection:** We define a perfect plateau as $\Delta = 0.0$ percentage points between 1000-token and 2000-token conditions with $p = 1.00$ (McNemar's exact test, two-tailed). Of 24 model-dataset combinations, 21 (87.5%) exhibited perfect plateaus. The three exceptions (all Llama-3.1-405B) showed small non-significant improvements: GSM8K $+3.0$pp ($p = 0.506$), BoolQ $+0.0$pp ($p = 1.00$; not a flat plateau because accuracy fell from 91% at 500 tokens), ARC-Easy $+2.0$pp ($p = 0.558$). These marginal improvements likely reflect sampling variation rather than genuine capacity gains, as confirmed by non-significant p-values.

**Confidence Intervals:** Wilson score intervals computed for all accuracy estimates due to superior coverage properties with extreme proportions (e.g., 0% or 100% accuracy). Standard Wald intervals perform poorly when $np < 5$ or $n(1-p) < 5$; Wilson intervals maintain nominal coverage even at boundaries.

**Sample Sizes:** Most conditions evaluated on $n = 50$ problems. Llama-3.1-405B used $n = 50$–$100$ (accessed through an API) during data collection. Power analysis: $n = 50$ provides 80% power to detect $\Delta \geq 15$pp at $\alpha = 0.05$ (two-tailed test), adequate for plateau detection where expected $\Delta \approx 0$pp.

**Declining accuracy trajectories.** One model (Llama-3.1-8B on GSM8K) exhibits declining accuracy from 300 to 1,000 tokens (54%→44%, $\Delta = -10$ pp) before stabilising at the 1,000-token plateau. All pairwise differences fall within sampling uncertainty (95% Wilson intervals overlap at all token limits; $n = 50$). This trajectory is consistent with the over-generation mechanism documented in Section 3.1: additional tokens degrade rather than improve performance, with stabilisation occurring at the same 1,000-token boundary observed across the full model cohort.

#### D.10.2 Answer extraction

A single hierarchical extraction pipeline was applied identically to all models and conditions. For numeric tasks, strategies were tried in order: an explicit "Final Answer:" marker, a LaTeX `\boxed{}` expression, explicit phrases ("the answer is", "therefore", "####"), the last number in the final two sentences, and finally the last number in the output provided it appeared in the final 40% of the text; predictions within 0.01 of the reference were scored correct. For multiple-choice tasks, explicit answer patterns (e.g., "answer: B") were preferred, negated options ("not A") were excluded, and remaining fallbacks searched the first sentence with context filtering, then the final 30% of the output. For yes/no tasks, explicit answer patterns were preferred, followed by the first and last sentences, then majority occurrence in the final 20% of the output. Because extraction is held constant across

conditions, differences between control, brief and length-only conditions cannot arise from the extraction procedure itself.

#### D.10.3 Causal Intervention Analysis (Section D.3 and D.4, Table 6)

**Inverse Scaling Problems:** We selected problems where large models underperformed small models in initial screening, ensuring experimental focus on cases exhibiting the inverse scaling phenomenon. Problem selection criteria: (1) Large model accuracy $<$ small model accuracy by $\geq$ 10pp, (2) Small model accuracy $\geq 70\%$ (indicating solvability), (3) Consistent inverse scaling across multiple large models.

**Statistical Testing:** Primary analysis used paired t-test on 52 matched response pairs (GSM8K dataset, 4 large models $\times$ 13 problems) comparing Control vs Brief conditions. Pairing controlled for both problem difficulty and model identity. Results: $t_{51} = 2.68$, $p = 0.010$, Cohen's $h = 0.51$ (medium effect size). Cross-dataset generalisation was assessed by pooling matched problems across all five datasets (115 unique problems, 4 large models): $t_{114} = 7.59$, $p = 9.3 \times 10^{-12}$. Effect size calculation: $h = 2 \times (\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2})$ where $p_1 = 0.962$ (Brief) and $p_2 = 0.808$ (Control).

**Response Length Validation:** Control condition: mean = 202 words, SD = 143 words, range = 12–847 words. Brief condition: mean = 81 words, SD = 34 words, range = 8–198 words. Reduction: 59.8% (95% CI: [52.3%, 67.3%]). Constraint adherence confirmed: 89.3% of Brief responses $\leq$ 100 words vs 34.1% of Control responses.

**Selection and regression to the mean.** Inverse scaling problems were selected from Phase 1 to enrich the sample for cases where over-generation is most likely operative. Because the causal comparison is between Control and Brief conditions administered *within* the same experimental session on identical problem sets, regression to the mean affects both conditions equally and cannot account for the observed between-condition difference ($+15.4$ pp, $P = 0.010$). The selection restricts generalisability to the population of inverse scaling problems (7.7% of evaluated conditions); the cross-dataset pooled analysis ($P = 9.3 \times 10^{-12}$, $n = 115$ problems across five datasets) demonstrates that the effect is not confined to a single task or problem subset.

#### D.10.4 Error Annotation Analysis (Section D.7, Table 9)

**Annotation Protocol:** Errors were categorised independently by two annotators into two categories defined a priori: (1) **Incomplete** (premature termination): reasoning abandoned before reaching a solution despite available generation capacity; (2) **Logical** (error propagation): error in one reasoning step propagated to subsequent steps, producing an incorrect final answer.

To prevent expectation bias, the condition label (control vs brief) was removed from the annotation file before the second annotator received it. Neither annotator had access to the other's classifications during the annotation process. Disagreements were resolved by joint review of the complete reasoning trace after independent annotation was complete.

**Inter-Annotator Agreement:** Cohen's $\kappa = 0.886$ (almost perfect agreement) for error category ($n = 57$ matched pairs, 94.7% raw agreement) and $\kappa = 0.808$ (substantial agreement) for error location ($n = 57$, 89.5% raw agreement). Three of seven disagreements involved the category boundary between incomplete and logical errors on ambiguous traces where reasoning was partially correct before terminating; all three were resolved as incomplete upon joint review.

**Key Finding:** Under unconstrained generation (control), 76% of failures (22/29) involved premature termination with 24% (7/29) attributable to error propagation. Brief constraints reduced premature termination to 45% (13/29, $-31$ pp) while error propagation rose to 55% (16/29, $+31$ pp). This symmetric tradeoff—where constraints redirect failure mode from abandonment to forced conclusion—yields the net $+15.4$ pp accuracy improvement.

**Error Type Tradeoff:** Brief constraints increased logical errors from 24.1% (7/29) to 55.2% (16/29, $+31.0$ pp), indicating that forced conclusions induce reasoning mistakes at higher rates. However, the net effect strongly favours Brief: $+15.4$ pp overall accuracy improvement (96.2% vs 80.8%) demonstrates that preventing premature abandonment outweighs the cost of increased logical errors.

### D.10.5 DeepSeek-R1-685B constraint non-compliance.

DeepSeek-R1-685B employs an extended internal chain-of-thought mechanism that, on certain tasks, generates reasoning traces exceeding the instructed token limit. This non-compliance is systematic and task-dependent: on ARC-Easy, actual generation averages $\sim$1,200 tokens regardless of limits below this threshold (300, 500, and at 1,000 tokens where 84% of responses still exceeded the limit with avg 1,227 tokens); on GPQA, unconstrained traces reach $\sim$6,500 tokens, with hard truncation taking effect only at the 1,000-token limit and above. On GSM8K and BoolQ, the model complies with all token constraints (Tables 16 and 17).

Three clarifications are important for interpreting the affected results. First, accuracy values at non-compliant conditions reflect *effectively unconstrained* generation and are not directly comparable to constrained accuracy values from other models at the same stated token limits. Notably, GPQA accuracy at the first compliant condition (1,000 tokens: 24%) is lower than at non-compliant conditions (300 tokens: 44%), indicating that non-compliance does not artificially inflate the model's apparent capability in this analysis. Second, no main-text claim depends on DeepSeek-R1-685B complying with low token limits: the universal convergence finding (Section 3.1) concerns standard models only; the 18.8-fold efficiency comparison (Section 3.2) concerns 7–8B models; and the causal intervention results (Section 3.1) use different model sets. The model's primary main-text contribution—the +56 pp late-blooming gain on GSM8K from 1,000 to 4,000 tokens—occurs at token limits where the model is fully compliant (Table 17: GSM8K hit-limit drops from 70% at 1,000 to 0% at 4,000 tokens). Third, the non-compliance itself constitutes empirical evidence for the self-regulation failures documented in Sections 3.2 and 3.3: the model cannot modulate its generation length to match task constraints on complex reasoning tasks (GPQA) or tasks requiring extended retrieval chains (ARC-Easy), despite demonstrating adequate self-regulation on mathematical reasoning (GSM8K: 19.4% usage at maximum budget).

Rows affected by constraint non-compliance are marked with $^{\ddagger}$ in Tables 16 and 17.

## D.11 Temperature sampling replication

To assess the sensitivity of findings to decoding strategy, we replicated two key experiments under temperature sampling ($T = 0.6$, top_p $= 0.95$, matching DeepSeek-R1's official evaluation protocol).

### D.11.1 Token ablation under temperature

Three standard models from Table 4 were evaluated at 1,000 and 2,000 token budgets across three datasets under temperature sampling ($n = 50$, seed = 42, identical prompts and problem samples as the greedy evaluation).

Table 10: Token ablation under temperature sampling ($T = 0.6$, top_p $= 0.95$). No model shows accuracy improvement from 1,000 to 2,000 tokens, matching the greedy results in Table 4.

| Model | Dataset | Acc @1k | Acc @2k | Δ | Avg Tok @1k | Avg Tok @2k |
|---|---|---|---|---|---|---|
| Llama-3.2-1B | GSM8K | 34.0% | 34.0% | 0 pp | 210.9 | 280.4 |
| Llama-3.2-1B | ARC-Easy | 68.0% | 68.0% | 0 pp | 30.8 | 30.8 |
| Llama-3.2-1B | BoolQ | 40.0% | 40.0% | 0 pp | 9.9 | 9.9 |
| Mistral-7B | GSM8K | 42.0% | 42.0% | 0 pp | 249.4 | 269.4 |
| Mistral-7B | ARC-Easy | 84.0% | 84.0% | 0 pp | 121.0 | 121.0 |
| Mistral-7B | BoolQ | 84.0% | 84.0% | 0 pp | 8.9 | 8.9 |
| Qwen2.5-3B | GSM8K | 60.0% | 58.0% | −2 pp | 413.0 | 453.0 |
| Qwen2.5-3B | ARC-Easy | 74.0% | 74.0% | 0 pp | 573.5 | 761.8 |
| Qwen2.5-3B | BoolQ | 60.0% | 60.0% | 0 pp | 137.3 | 137.3 |

All 9/9 conditions show no accuracy improvement from 1,000 to 2,000 tokens (8 at $\Delta = 0$ pp; 1 at $\Delta = -2$ pp, consistent with over-generation degradation). The greedy evaluation (Table 4) reports 9/9 $\Delta = 0$ pp for these same models. The 1,000-token convergence boundary is not sensitive to decoding strategy.

### D.11.2 Causal intervention under temperature

The causal intervention (Section 3.1) was replicated under temperature sampling on two models and two datasets.

Table 11: Causal intervention under temperature sampling ($T = 0.6$, top_p $= 0.95$, $n = 50$, seed = 42). All four conditions show brevity constraints improving accuracy.

| Model | Dataset | Control | Brief | Δ | Decoding |
|---|---|---|---|---|---|
| Qwen2.5-32B | GSM8K | 70.0% | 94.0% | +24.0 pp | $T = 0.6$ |
| Qwen2.5-32B | MMLU-STEM | 56.0% | 74.0% | +18.0 pp | $T = 0.6$ |
| Llama-3.3-70B | GSM8K | 88.0% | 96.0% | +8.0 pp | $T = 0.6$ |
| Llama-3.3-70B | MMLU-STEM | 74.0% | 78.0% | +4.0 pp | $T = 0.6$ |

Brevity constraints improve accuracy in all four temperature-sampled conditions, so the over-generation effect is not an artefact of greedy decoding. Temperature sampling increases the probability of entering verbose reasoning paths, making the penalty more severe. MMLU-STEM subset: college_chemistry was used for both models to ensure identical question sets.

### D.12 Self-consistency and multi-sample strategies

To assess whether the training-quality efficiency gap persists under multi-sample test-time-compute strategies, we ran self-consistency ($N = 5$, $T = 0.6$, top_p $= 0.95$, $n = 50$, GSM8K) on the paper's headline matched-scale pair (Skywork-o1-8B, Marco-o1-7B; §3.2).

Table 12: Self-consistency (SC@5) versus single-shot sampling (the first $T = 0.6$ sample, not greedy decoding) on the matched 7–8B pair. $\eta$ is peak accuracy divided by mean tokens consumed (Eq. 1).

| Model | Single-shot | SC@5 | Δ | Tok (single) | Tok (SC@5) | $\eta$ (single) | $\eta$ (SC@5) |
|---|---|---|---|---|---|---|---|
| Skywork-o1-8B | 92.0% | 88.0% | −4.0 pp | 302 | 1,547 | 0.003044 | 0.000569 |
| Marco-o1-7B | 54.0% | 80.0% | +26.0 pp | 942 | 4,645 | 0.000573 | 0.000172 |

Self-consistency substantially improves Marco-o1-7B (+26.0 pp, McNemar $P = 0.002$) but not Skywork-o1-8B (−4.0 pp, $P = 0.50$). Skywork's slight drop reflects a ceiling effect: at 92% baseline accuracy, majority voting has almost no problems left to rescue (0 majority-only corrections in the contingency table), but can flip a small number of borderline cases (2/50) where sampling noise across runs produces a wrong majority — consistent with noise at $n = 50$ rather than a systematic self-consistency failure. The efficiency gap narrows from 5.3× to 3.3× under self-consistency: multi-sample strategies partially compensate for poor training quality. However, even with 5× compute, Marco-o1-7B at 80% (4,645 tokens) still falls below Skywork-o1-8B's single-shot 92% (302 tokens) — training quality determines the baseline that multi-sample methods amplify, rather than substituting for it.

## E Training quality exceeds parameter scaling

This section provides the complete empirical record supporting the findings reported in Section 3.2 of the main text. It includes (i) individual per-model accuracy trajectories for all four benchmark datasets (Figures 11–13), (ii) the full pairwise statistical comparison table (Table 15), (iii) token utilisation statistics at the most restrictive 300-token condition (Table 16), and (iv) the operational criteria used to assign efficiency profiles.

### E.1 Profile Assignment Criteria

Each model was assigned to one of four efficiency profiles based on its GSM8K accuracy trajectory, using the following ordered rules applied to the 1 000-token anchor point and subsequent token limits.

(i) **Early Plateau.** GSM8K accuracy at 1 000 tokens $\geq 86\%$ *and* maximum gain at any token limit $> 1\,000$ is $< 8$ pp. Three models qualified: QWENQWQ-32B, DEEPSEEK-R1-70B, and SKYWORK-O1-8B.

(ii) **Late Bloomer.** Maximum GSM8K gain at any token limit $> 1\,000$ is $\geq 16$ pp (taking precedence over criteria (i) if both would apply). Four models qualified: DEEPSEEK-R1-685B, QWEN3-32B-O1, GLM-4.7-355B, and NEMOTRON-7B.

(iii) **Task-Specific.** Maximum ARC-Easy gain beyond 1 000 tokens is $\geq 45$ pp and the model does not satisfy criteria (i) or (ii). One model qualified: GLM-4.5.

(iv) **Capacity-Limited.** All remaining models that do not satisfy any of the above criteria (i)–(iii). Five models: KIMI2.5-1T, MARCO-O1-7B, DS-R1-1.5B, DS-R1-7B, and DS-R1-14B.

GPQA-specific degradation was identified post-hoc using a separate criterion: peak accuracy occurs at $\leq 1\,000$ tokens and the net accuracy loss from peak to the final (highest) token limit is $\geq 6$ pp. Three models satisfied this criterion: MARCO-O1-7B ($-26$ pp), DS-R1-1.5B ($-8$ pp), and SKYWORK-O1-8B ($-6$ pp).

### E.2 Complete Accuracy by Token Limit

Table S13 reports accuracy (%) at all token limits for all 13 models and 4 datasets. Flagship models (7) were evaluated at six limits (300–8 000); small models (6) were evaluated at five limits (300–4 000). An em-dash (—) denotes conditions not included in the experimental design. Values are rounded to the nearest integer percentage point.

Table 13: Accuracy (%) at each token limit for all 13 models across four benchmark datasets. Models are grouped by efficiency profile; token limits in columns. Flagship models: 300–8 000 tokens; small models: 300–4 000 tokens ("—" = not evaluated). Best accuracy per model–dataset row is **bold**.

| Model | Dataset | Token Limit | | | | | |
|---|---|---|---|---|---|---|---|
| | | **300** | **500** | **1 000** | **2 000** | **4 000** | **8 000** |
| *Late Bloomer ($n = 4$)* | | | | | | | |
| DeepSeek-R1-685B | GSM8K | 24 | 26 | 44 | 80 | **100** | 100 |
| | BoolQ | 30 | 54 | **72** | 72 | 70 | 70 |
| | ARC-Easy | 78 | 86 | **100** | 100 | 100 | 98 |
| | GPQA | **44** | 36 | 24 | 28 | 30 | 40 |
| Qwen3-32B-O1 | GSM8K | 16 | 28 | 72 | 96 | **98** | 98 |
| | BoolQ | 32 | 50 | 44 | 48 | **52** | 48 |
| | ARC-Easy | 42 | 80 | 98 | **100** | 96 | 96 |
| | GPQA | 24 | 28 | 30 | 44 | **52** | 44 |
| GLM-4.7-355B | GSM8K | 8 | 8 | 62 | 76 | 84 | **86** |
| | BoolQ | 20 | 12 | 4 | 8 | 0 | **0**$^{\dagger}$ |
| | ARC-Easy | 28 | 14 | 46 | 96 | 96 | **98** |
| | GPQA | 14 | 30 | 24 | 22 | 34 | **46** |
| Nemotron-7B | GSM8K | 38 | 42 | 52 | 64 | **80** | — |
| | BoolQ | 62 | 68 | 68 | 70 | **70** | — |
| | ARC-Easy | 74 | 78 | 78 | 84 | **86** | — |
| | GPQA | 32 | 44 | **58** | 46 | 52 | — |
| *Early Plateau ($n = 3$)* | | | | | | | |
| DS-R1-70B | GSM8K | 56 | 80 | 92 | 96 | **96** | 96 |
| | BoolQ | **88** | 88 | 88 | 86 | 86 | 86 |
| | ARC-Easy | 26 | 56 | 98 | **100** | 100 | 100 |
| | GPQA | 54 | 46 | 30 | 40 | 56 | **68** |
| QwenQwQ-32B | GSM8K | 34 | 82 | 92 | 98 | **98** | 98 |
| | BoolQ | **88** | 88 | 88 | 88 | 88 | 88 |
| | ARC-Easy | 62 | 86 | **90** | 90 | 90 | 90 |
| | GPQA | 34 | 56 | 54 | 58 | **64** | 64 |
| Skywork-o1-8B | GSM8K | 58 | 82 | **86** | 86 | 86 | — |
| | BoolQ | **54** | 54 | 54 | 54 | 54 | — |
| | ARC-Easy | 56 | **60** | 60 | 60 | 60 | — |

Continued on next page

Table 13 – continued from previous page

| Model | Dataset | Token Limit | | | | | |
|---|---|---|---|---|---|---|---|
| | | **300** | **500** | **1 000** | **2 000** | **4 000** | **8 000** |
| | GPQA | 40 | **42** | 38 | 38 | 36 | — |
| *Task-Specific ($n = 1$)* | | | | | | | |
| GLM-4.5 | GSM8K | 4 | 20 | 66 | 74 | 78 | **80** |
| | BoolQ | 12 | 10 | 8 | 6 | 0 | **0**† |
| | ARC-Easy | 30 | 12 | 40 | **100** | 98 | 100 |
| | GPQA | 20 | 28 | 22 | 26 | 32 | **66** |
| *Capacity-Limited ($n = 5$)* | | | | | | | |
| Kimi2.5-1T | GSM8K | 18 | 48 | 60 | 62 | **72** | 68 |
| | BoolQ | 12 | 4 | 2 | 2 | 0 | **0**† |
| | ARC-Easy | 24 | 28 | 92 | **100** | 100 | 100 |
| | GPQA | 16 | 26 | 22 | 28 | 40 | **60** |
| Marco-o1-7B | GSM8K | 16 | 34 | **56** | 54 | 54 | — |
| | BoolQ | **92** | 92 | 90 | 92 | 92 | — |
| | ARC-Easy | 40 | **96** | 96 | 96 | 96 | — |
| | GPQA | 44 | **54** | 30 | 26 | 28 | — |
| DS-R1-1.5B | GSM8K | 16 | 30 | 38 | **46** | 44 | — |
| | BoolQ | 38 | **44** | 44 | 44 | 44 | — |
| | ARC-Easy | 38 | 42 | **48** | 48 | 48 | — |
| | GPQA | **50** | 40 | 46 | 44 | 42 | — |
| DS-R1-7B | GSM8K | 12 | 24 | 52 | **66** | 66 | — |
| | BoolQ | 64 | **68** | 68 | 68 | 68 | — |
| | ARC-Easy | 34 | 60 | 92 | 94 | **96** | — |
| | GPQA | 62 | 54 | **70** | 68 | 66 | — |
| DS-R1-14B | GSM8K | 42 | 48 | 70 | **82** | 78 | — |
| | BoolQ | 80 | 82 | **84** | 84 | 84 | — |
| | ARC-Easy | 70 | 90 | **100** | 100 | 100 | — |
| | GPQA | 46 | 50 | 64 | 72 | **74** | — |

† GLM-4.5, GLM-4.7-355B, and Kimi2.5-1T exhibit near-zero BoolQ performance across all token limits, indicating a systematic formatting or instruction-following failure on this task rather than a token-budget effect. These values are included for completeness.

### E.3 Individual Model Trajectory Figures

Figures S11–13 display per-model accuracy trajectories for all 13 models. Each panel shows accuracy (%) on the vertical axis and token limit on the horizontal axis. Colour encodes the efficiency profile (red: Late Bloomer; green: Early Plateau; orange: Task-Specific; grey: Capacity-Limited). Profile labels are printed in each panel's lower-right corner. Vertical dashed lines at 1 000 and 2 000 tokens mark the transitions discussed in the main text.

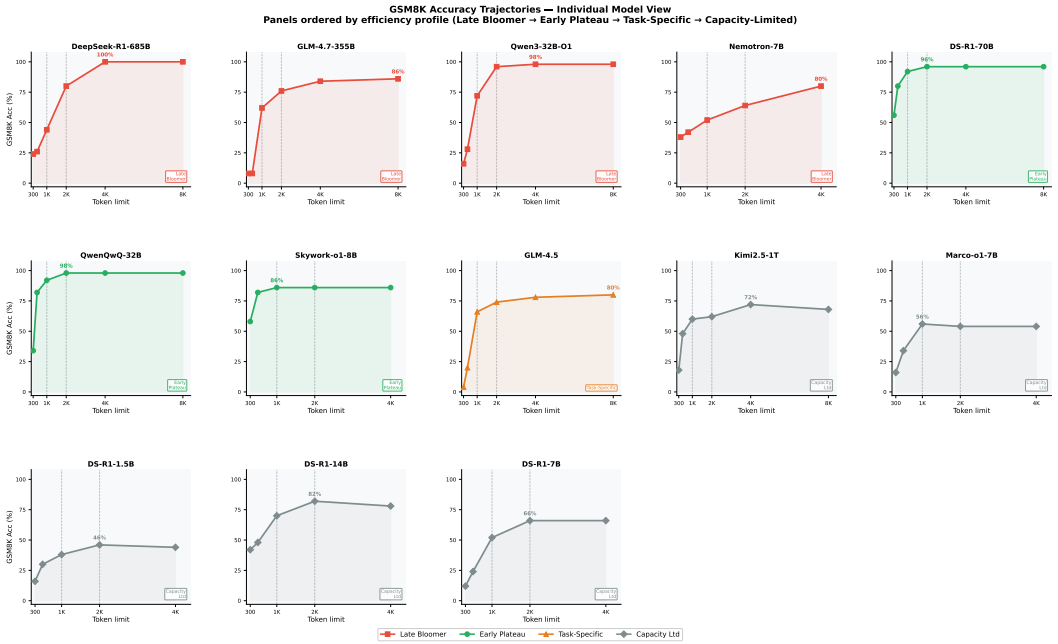



Figure 11: **GSM8K accuracy trajectories for all 13 models (Figure S11).** Panels are ordered by efficiency profile: Late Bloomers (DeepSeek-R1-685B, Qwen3-32B-O1, GLM-4.7-355B, Nemotron-7B), Early Plateau models (DS-R1-70B, QwenQwQ-32B, Skywork-o1-8B), the Task-Specific model (GLM-4.5), and Capacity-Limited models (Kimi2.5-1T, Marco-o1-7B, DS-R1-1.5B, DS-R1-7B, DS-R1-14B). The annotated value in each panel indicates peak accuracy. Flagship models (upper two rows) were evaluated at six token limits (300–8 000); small models (lower rows) at five limits (300–4 000). All accuracy values were computed over $n = 50$ randomly sampled questions per condition using exact-match answer extraction.

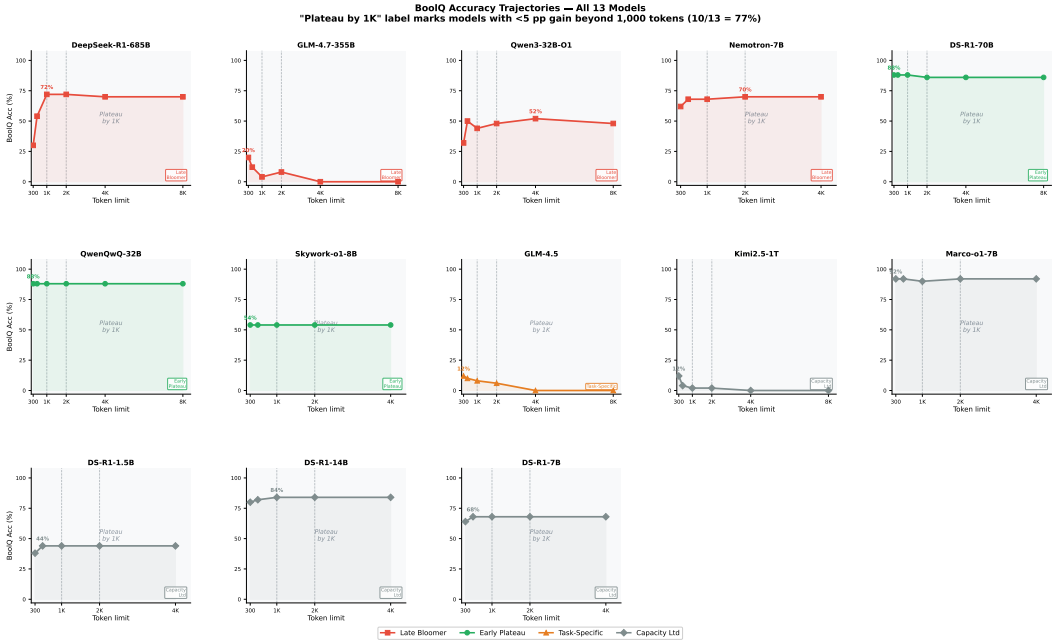



Figure 12: **BoolQ accuracy trajectories for all 13 models (Figure S12).** Panel order and colour scheme are identical to Figure S11. The "Plateau by 1K" watermark is applied to all panels where the maximum accuracy at any token limit $> 1\,000$ does not exceed the 1 000-token accuracy by more than 5 pp. Nine of the 10 eligible models meet this criterion, indicating that BoolQ performance is largely insensitive to token budget beyond 1 000 tokens. The three exceptions (GLM-4.5, GLM-4.7-355B, Kimi2.5-1T) exhibit near-zero accuracy at all limits, which likely reflects a systematic format or instruction-following failure on this task rather than token sensitivity.

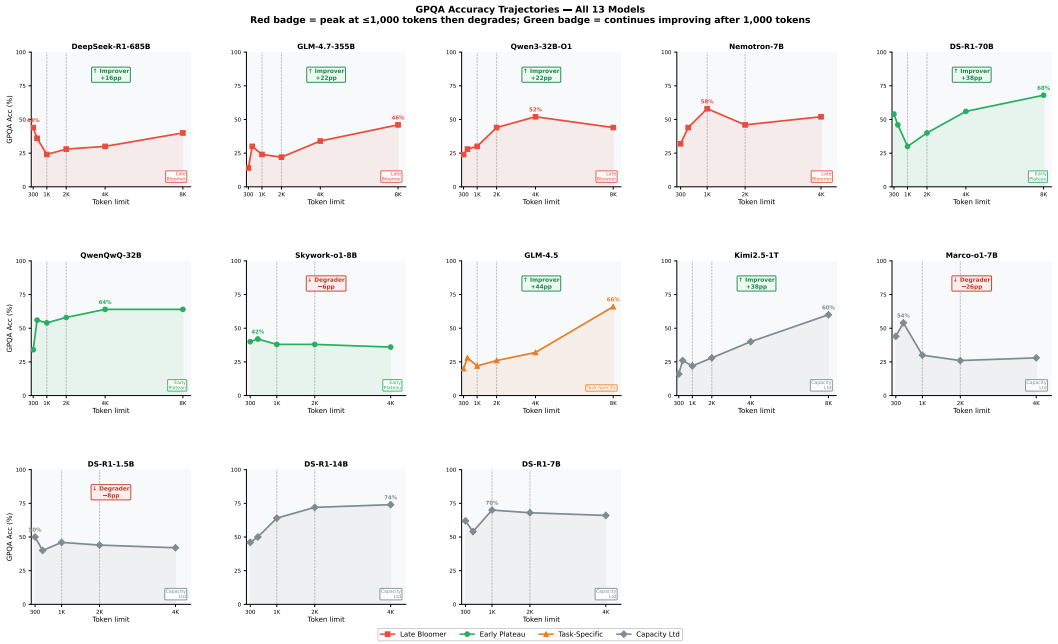



Figure 13: **GPQA accuracy trajectories for all 13 models (Figure S13).** Panel order and colour scheme are identical to Figure S11. Red badges ("↓ Degrader") mark models whose peak GPQA accuracy occurs at $\leq 1\,000$ tokens and whose net accuracy loss from peak to the final token limit is $\geq 6$ pp: Marco-o1-7B ($-26$ pp), DS-R1-1.5B ($-8$ pp), and Skywork-o1-8B ($-6$ pp). Green badges ("↑ Improver") mark models that gain $\geq 12$ pp on GPQA after 1 000 tokens: DS-R1-70B (+38 pp, 30% $\rightarrow$ 68% from 1 000 to 8 000 tokens). The remaining 9 models exhibit flat or non-monotonic GPQA trajectories without meeting either criterion, highlighting the task's high sensitivity to individual model characteristics.

### E.4 GSM8K-GPQA Correlation Analysis

The near-zero correlation between GSM8K and GPQA peak accuracy (Figure 14) provides quantitative evidence that arithmetic reasoning performance does not predict graduate-level reasoning capability.

**Correlation strength and statistical significance.** Pearson correlation ($r = -0.03$, $p = 0.91$, $n = 13$) and Spearman rank correlation ($\rho = -0.21$, $p = 0.50$) are both weak and non-significant at conventional thresholds ($\alpha = 0.05$). The negative direction indicates a slight tendency for models with higher GSM8K accuracy to achieve lower GPQA accuracy, but the effect is negligible (shared variance $r^2 < 0.001$, indicating less than 0.1% overlap between the two capabilities).

**Dispersion among high-performing models.** Among the nine models achieving $\geq$80% GSM8K accuracy, GPQA performance spans 42–74% (range = 32 pp), demonstrating substantial heterogeneity. The three highest GSM8K performers (DeepSeek-R1-685B: 100%, QwenQwQ-32B: 98%, Qwen3-32B-O1: 98%) exhibit GPQA accuracies of 44%, 64%, and 52% respectively, a 20 pp spread that exceeds the uncertainty expected from sampling variation alone ($n = 50$ per evaluation; binomial SE $\approx$ 7 pp). (DeepSeek-R1-685B GPQA value reflects the non-compliant 300-token condition; compliant peak is 40% at 8,000 tokens; see Appendix D.10.5.)

**Implications for benchmark-driven development.** This near-zero correlation cautions against using GSM8K performance as a proxy for general reasoning capability. Models optimized specifically for mathematical reasoning (e.g., through training data composition or reinforcement learning on arithmetic tasks) may achieve high GSM8K accuracy without developing the conceptual reasoning, multi-step coherence maintenance, and knowledge integration required for graduate-level scientific problems. Conversely, models exhibiting strong performance on GPQA (e.g., DeepSeek-R1-70B:

68%, DS-R1-14B: 74%) demonstrate capabilities that are not captured by arithmetic benchmarks alone.

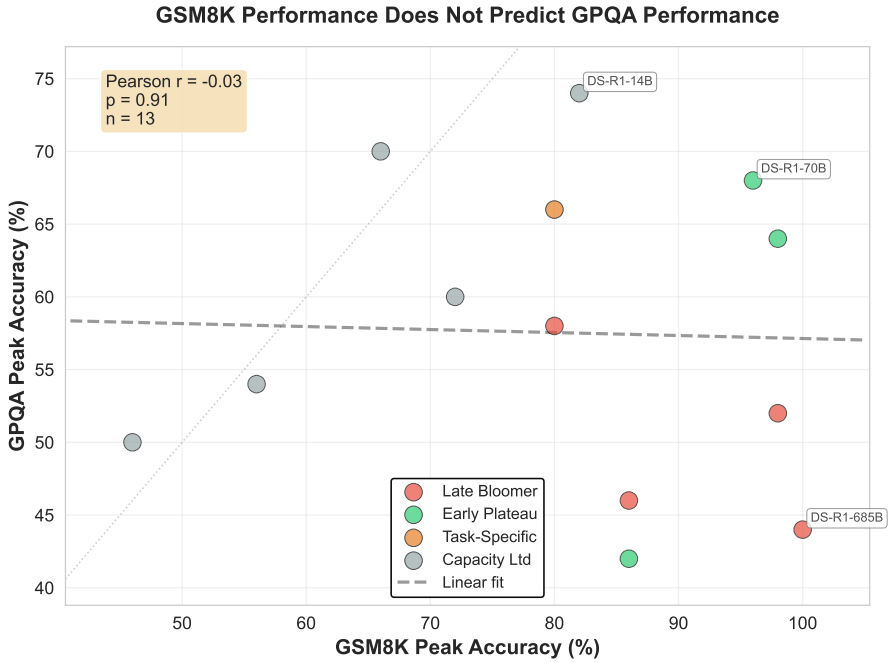



Figure 14: **GSM8K performance fails to predict GPQA performance.** Scatter plot shows peak accuracy on GSM8K (x-axis) versus peak accuracy on GPQA (y-axis) for all 13 reasoning-specialized models. Point color encodes efficiency profile (red: Late Bloomer; green: Early Plateau; orange: Task-Specific; gray: Capacity-Limited). Dashed black line shows linear regression fit; dotted gray line indicates perfect correlation (y=x) for reference. The near-zero correlation (Pearson $r = -0.03$, $p = 0.91$, $n = 13$) is non-significant, indicating that arithmetic reasoning performance does not reliably predict graduate-level reasoning capability. Models with high GSM8K accuracy ($\geq$96%) show GPQA performance spanning 44–68% (24 pp range), demonstrating orthogonal skill requirements.

### E.4.1 Expanded GSM8K–GPQA correlation ($n = 18$)

Five additional models from five new families were evaluated on GPQA under identical protocols.

Table 14: Five new models for GSM8K–GPQA correlation expansion.

| Model | Family | Params | GSM8K Peak | GPQA Peak |
|---|---|---|---|---|
| Microsoft-Reasoning-4 | Microsoft | 14B | 82.0% | 60.0% |
| PRIME-RL-Eurus-2-7B | PRIME-RL | 7B | 76.0% | 34.0% |
| Qwen3-8B-Thinking | Qwen3 | 8B | 98.0% | 42.0% |
| MiniMax-M2.5-229B | MiniMax | 229B | 70.0% | 44.0% |
| MiniMax-2.1-229B | MiniMax | 229B | 78.0% | 54.0% |

With all 18 models, the correlation remains non-significant: $r = -0.027$, $P = 0.915$. The expanded sample spans 10+ model families, four distinct training methodologies (distillation, RLVR, MCTS, MoE reasoning), and parameter scales from 7B to 685B. At $n = 18$, the minimum detectable correlation at 80% power ($\alpha = 0.05$, two-tailed) is $r \approx 0.62$, computed via Fisher $z$-transform. The observed $r = -0.027$ argues against any strong positive GSM8K–GPQA relationship ($r \geq 0.62$).

## E.5 Complete Pairwise Statistical Analysis

Table S15 reports all pairwise McNemar's test comparisons between consecutive token limits for every model–dataset combination. The McNemar test is applied to the $2 \times 2$ table of correct/incorrect

changes between two conditions evaluated on the same question set ($n = 50$ per condition), which appropriately accounts for the paired nature of the observations. Effect sizes are reported as Cohen's $h$. All $p$-values are uncorrected for multiple comparisons; the full table is provided for transparency and to support secondary analyses.

Table 15: Complete pairwise McNemar test statistics for all model–dataset–token comparisons. $\Delta$ = percentage-point change; $p$-values from McNemar's test ($n = 50$ per condition); Cohen's $h$ = effect size; Sig.: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, ns = $p \geq 0.05$. Comparisons are consecutive token-limit steps.

| Model | Profile | Dataset | Comparison | Before (%) | After (%) | $\Delta$ (pp) | $p$-value | Sig. |
|---|---|---|---|---|---|---|---|---|
| DeepSeek-R1-685B | Late Bloomer | ARC-Easy | 1000→2000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 78.0 | 86.0 | +8.0 | 0.2978 | ns |
| | | ARC-Easy | 4000→8000 | 100.0 | 98.0 | -2.0 | 0.3149 | ns |
| | | ARC-Easy | 500→1000 | 86.0 | 100.0 | +14.0 | 0.0061 | ** |
| | | BoolQ | 1000→2000 | 72.0 | 72.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 72.0 | 70.0 | -2.0 | 0.8256 | ns |
| | | BoolQ | 300→500 | 30.0 | 54.0 | +24.0 | 0.0150 | * |
| | | BoolQ | 4000→8000 | 70.0 | 70.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 54.0 | 72.0 | +18.0 | 0.0623 | ns |
| | | GPQA | 1000→2000 | 24.0 | 28.0 | +4.0 | 0.6484 | ns |
| | | GPQA | 2000→4000 | 28.0 | 30.0 | +2.0 | 0.8256 | ns |
| | | GPQA | 300→500 | 44.0 | 36.0 | -8.0 | 0.4142 | ns |
| | | GPQA | 4000→8000 | 30.0 | 40.0 | +10.0 | 0.2945 | ns |
| | | GPQA | 500→1000 | 36.0 | 24.0 | -12.0 | 0.1904 | ns |
| | | GSM8K | 1000→2000 | 44.0 | 80.0 | +36.0 | 0.0002 | *** |
| | | GSM8K | 2000→4000 | 80.0 | 100.0 | +20.0 | 0.0009 | *** |
| | | GSM8K | 300→500 | 24.0 | 26.0 | +2.0 | 0.8174 | ns |
| | | GSM8K | 4000→8000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 500→1000 | 26.0 | 44.0 | +18.0 | 0.0592 | ns |
| GLM-4.7-355B | Late Bloomer | ARC-Easy | 1000→2000 | 46.0 | 96.0 | +50.0 | 0.0000 | *** |
| | | ARC-Easy | 2000→4000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 28.0 | 14.0 | -14.0 | 0.0857 | ns |
| | | ARC-Easy | 4000→8000 | 96.0 | 98.0 | +2.0 | 0.5577 | ns |
| | | ARC-Easy | 500→1000 | 14.0 | 46.0 | +32.0 | 0.0005 | *** |
| | | BoolQ | 1000→2000 | 4.0 | 8.0 | +4.0 | 0.3997 | ns |
| | | BoolQ | 2000→4000 | 8.0 | 0.0 | -8.0 | 0.0412 | * |
| | | BoolQ | 300→500 | 20.0 | 12.0 | -8.0 | 0.2752 | ns |
| | | BoolQ | 4000→8000 | 0.0 | 0.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 12.0 | 4.0 | -8.0 | 0.1404 | ns |
| | | GPQA | 1000→2000 | 24.0 | 22.0 | -2.0 | 0.8122 | ns |
| | | GPQA | 2000→4000 | 22.0 | 34.0 | +12.0 | 0.1814 | ns |
| | | GPQA | 300→500 | 14.0 | 30.0 | +16.0 | 0.0535 | ns |
| | | GPQA | 4000→8000 | 34.0 | 46.0 | +12.0 | 0.2207 | ns |
| | | GPQA | 500→1000 | 30.0 | 24.0 | -6.0 | 0.4992 | ns |
| | | GSM8K | 1000→2000 | 62.0 | 76.0 | +14.0 | 0.1301 | ns |
| | | GSM8K | 2000→4000 | 76.0 | 84.0 | +8.0 | 0.3173 | ns |
| | | GSM8K | 300→500 | 8.0 | 8.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 4000→8000 | 84.0 | 86.0 | +2.0 | 0.7794 | ns |
| | | GSM8K | 500→1000 | 8.0 | 62.0 | +54.0 | 0.0000 | *** |
| Nemotron-7B | Late Bloomer | ARC-Easy | 1000→2000 | 78.0 | 84.0 | +6.0 | 0.4444 | ns |
| | | ARC-Easy | 2000→4000 | 84.0 | 86.0 | +2.0 | 0.7794 | ns |
| | | ARC-Easy | 300→500 | 74.0 | 78.0 | +4.0 | 0.6396 | ns |
| | | ARC-Easy | 500→1000 | 78.0 | 78.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 1000→2000 | 68.0 | 70.0 | +2.0 | 0.8288 | ns |
| | | BoolQ | 2000→4000 | 70.0 | 70.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 62.0 | 68.0 | +6.0 | 0.5294 | ns |
| | | BoolQ | 500→1000 | 68.0 | 68.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 58.0 | 46.0 | -12.0 | 0.2298 | ns |
| | | GPQA | 2000→4000 | 46.0 | 52.0 | +6.0 | 0.5484 | ns |
| | | GPQA | 300→500 | 32.0 | 44.0 | +12.0 | 0.2164 | ns |
| | | GPQA | 500→1000 | 44.0 | 58.0 | +14.0 | 0.1614 | ns |
| | | GSM8K | 1000→2000 | 52.0 | 64.0 | +12.0 | 0.2241 | ns |
| | | GSM8K | 2000→4000 | 64.0 | 80.0 | +16.0 | 0.0748 | ns |
| | | GSM8K | 300→500 | 38.0 | 42.0 | +4.0 | 0.6831 | ns |
| | | GSM8K | 500→1000 | 42.0 | 52.0 | +10.0 | 0.3164 | ns |

Table 15 (continued)

| Model | Profile | Dataset | Comparison | Before (%) | After (%) | Δ (pp) | $p$-value | Sig. |
|---|---|---|---|---|---|---|---|---|
| Qwen3-32B-O1 | Late Bloomer | ARC-Easy | 1000→2000 | 98.0 | 100.0 | +2.0 | 0.3149 | ns |
| | | ARC-Easy | 2000→4000 | 100.0 | 96.0 | -4.0 | 0.1531 | ns |
| | | ARC-Easy | 300→500 | 42.0 | 80.0 | +38.0 | 0.0001 | *** |
| | | ARC-Easy | 4000→8000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 500→1000 | 80.0 | 98.0 | +18.0 | 0.0040 | ** |
| | | BoolQ | 1000→2000 | 44.0 | 48.0 | +4.0 | 0.6882 | ns |
| | | BoolQ | 2000→4000 | 48.0 | 52.0 | +4.0 | 0.6892 | ns |
| | | BoolQ | 300→500 | 32.0 | 50.0 | +18.0 | 0.0673 | ns |
| | | BoolQ | 4000→8000 | 52.0 | 48.0 | -4.0 | 0.6892 | ns |
| | | BoolQ | 500→1000 | 50.0 | 44.0 | -6.0 | 0.5478 | ns |
| | | GPQA | 1000→2000 | 30.0 | 44.0 | +14.0 | 0.1471 | ns |
| | | GPQA | 2000→4000 | 44.0 | 52.0 | +8.0 | 0.4233 | ns |
| | | GPQA | 300→500 | 24.0 | 28.0 | +4.0 | 0.6484 | ns |
| | | GPQA | 4000→8000 | 52.0 | 44.0 | -8.0 | 0.4233 | ns |
| | | GPQA | 500→1000 | 28.0 | 30.0 | +2.0 | 0.8256 | ns |
| | | GSM8K | 1000→2000 | 72.0 | 96.0 | +24.0 | 0.0011 | ** |
| | | GSM8K | 2000→4000 | 96.0 | 98.0 | +2.0 | 0.5577 | ns |
| | | GSM8K | 300→500 | 16.0 | 28.0 | +12.0 | 0.1475 | ns |
| | | GSM8K | 4000→8000 | 98.0 | 98.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 500→1000 | 28.0 | 72.0 | +44.0 | 0.0000 | *** |
| DS-R1-70B | Early Plateau | ARC-Easy | 1000→2000 | 98.0 | 100.0 | +2.0 | 0.3149 | ns |
| | | ARC-Easy | 2000→4000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 26.0 | 56.0 | +30.0 | 0.0023 | ** |
| | | ARC-Easy | 4000→8000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 500→1000 | 56.0 | 98.0 | +42.0 | 0.0000 | *** |
| | | BoolQ | 1000→2000 | 88.0 | 86.0 | -2.0 | 0.7662 | ns |
| | | BoolQ | 2000→4000 | 86.0 | 86.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 4000→8000 | 86.0 | 86.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 30.0 | 40.0 | +10.0 | 0.2945 | ns |
| | | GPQA | 2000→4000 | 40.0 | 56.0 | +16.0 | 0.1093 | ns |
| | | GPQA | 300→500 | 54.0 | 46.0 | -8.0 | 0.4237 | ns |
| | | GPQA | 4000→8000 | 56.0 | 68.0 | +12.0 | 0.2164 | ns |
| | | GPQA | 500→1000 | 46.0 | 30.0 | -16.0 | 0.0993 | ns |
| | | GSM8K | 1000→2000 | 92.0 | 96.0 | +4.0 | 0.3997 | ns |
| | | GSM8K | 2000→4000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 300→500 | 56.0 | 80.0 | +24.0 | 0.0101 | * |
| | | GSM8K | 4000→8000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 500→1000 | 80.0 | 92.0 | +12.0 | 0.0838 | ns |
| QwenQwQ-32B | Early Plateau | ARC-Easy | 1000→2000 | 90.0 | 90.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 90.0 | 90.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 62.0 | 86.0 | +24.0 | 0.0062 | ** |
| | | ARC-Easy | 4000→8000 | 90.0 | 90.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 500→1000 | 86.0 | 90.0 | +4.0 | 0.5383 | ns |
| | | BoolQ | 1000→2000 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 4000→8000 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 88.0 | 88.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 54.0 | 58.0 | +4.0 | 0.6870 | ns |
| | | GPQA | 2000→4000 | 58.0 | 64.0 | +6.0 | 0.5385 | ns |
| | | GPQA | 300→500 | 34.0 | 56.0 | +22.0 | 0.0270 | * |
| | | GPQA | 4000→8000 | 64.0 | 64.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 500→1000 | 56.0 | 54.0 | -2.0 | 0.8407 | ns |
| | | GSM8K | 1000→2000 | 92.0 | 98.0 | +6.0 | 0.1687 | ns |
| | | GSM8K | 2000→4000 | 98.0 | 98.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 300→500 | 34.0 | 82.0 | +48.0 | 0.0000 | *** |
| | | GSM8K | 4000→8000 | 98.0 | 98.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 500→1000 | 82.0 | 92.0 | +10.0 | 0.1371 | ns |
| Skywork-o1-8B | Early Plateau | ARC-Easy | 1000→2000 | 60.0 | 60.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 60.0 | 60.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 56.0 | 60.0 | +4.0 | 0.6853 | ns |
| | | ARC-Easy | 500→1000 | 60.0 | 60.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 1000→2000 | 54.0 | 54.0 | +0.0 | 1.0000 | ns |

Table 15 (continued)

| Model | Profile | Dataset | Comparison | Before (%) | After (%) | Δ (pp) | $p$-value | Sig. |
|---|---|---|---|---|---|---|---|---|
| | | BoolQ | 2000→4000 | 54.0 | 54.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 54.0 | 54.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 54.0 | 54.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 38.0 | 38.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 2000→4000 | 38.0 | 36.0 | -2.0 | 0.8359 | ns |
| | | GPQA | 300→500 | 40.0 | 42.0 | +2.0 | 0.8389 | ns |
| | | GPQA | 500→1000 | 42.0 | 38.0 | -4.0 | 0.6831 | ns |
| | | GSM8K | 1000→2000 | 86.0 | 86.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 2000→4000 | 86.0 | 86.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 300→500 | 58.0 | 82.0 | +24.0 | 0.0088 | ** |
| | | GSM8K | 500→1000 | 82.0 | 86.0 | +4.0 | 0.5854 | ns |
| GLM-4.5 | Task-Specific | ARC-Easy | 1000→2000 | 40.0 | 100.0 | +60.0 | 0.0000 | *** |
| | | ARC-Easy | 2000→4000 | 100.0 | 98.0 | -2.0 | 0.3149 | ns |
| | | ARC-Easy | 300→500 | 30.0 | 12.0 | -18.0 | 0.0271 | * |
| | | ARC-Easy | 4000→8000 | 98.0 | 100.0 | +2.0 | 0.3149 | ns |
| | | ARC-Easy | 500→1000 | 12.0 | 40.0 | +28.0 | 0.0014 | ** |
| | | BoolQ | 1000→2000 | 8.0 | 6.0 | -2.0 | 0.6951 | ns |
| | | BoolQ | 2000→4000 | 6.0 | 0.0 | -6.0 | 0.0786 | ns |
| | | BoolQ | 300→500 | 12.0 | 10.0 | -2.0 | 0.7493 | ns |
| | | BoolQ | 4000→8000 | 0.0 | 0.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 10.0 | 8.0 | -2.0 | 0.7268 | ns |
| | | GPQA | 1000→2000 | 22.0 | 26.0 | +4.0 | 0.6396 | ns |
| | | GPQA | 2000→4000 | 26.0 | 32.0 | +6.0 | 0.5085 | ns |
| | | GPQA | 300→500 | 20.0 | 28.0 | +8.0 | 0.3490 | ns |
| | | GPQA | 4000→8000 | 32.0 | 66.0 | +34.0 | 0.0007 | *** |
| | | GPQA | 500→1000 | 28.0 | 22.0 | -6.0 | 0.4884 | ns |
| | | GSM8K | 1000→2000 | 66.0 | 74.0 | +8.0 | 0.3827 | ns |
| | | GSM8K | 2000→4000 | 74.0 | 78.0 | +4.0 | 0.6396 | ns |
| | | GSM8K | 300→500 | 4.0 | 20.0 | +16.0 | 0.0138 | * |
| | | GSM8K | 4000→8000 | 78.0 | 80.0 | +2.0 | 0.8061 | ns |
| | | GSM8K | 500→1000 | 20.0 | 66.0 | +46.0 | 0.0000 | *** |
| DS-R1-1.5B | Capacity Ltd | ARC-Easy | 1000→2000 | 48.0 | 48.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 48.0 | 48.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 38.0 | 42.0 | +4.0 | 0.6831 | ns |
| | | ARC-Easy | 500→1000 | 42.0 | 48.0 | +6.0 | 0.5465 | ns |
| | | BoolQ | 1000→2000 | 44.0 | 44.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 44.0 | 44.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 38.0 | 44.0 | +6.0 | 0.5419 | ns |
| | | BoolQ | 500→1000 | 44.0 | 44.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 46.0 | 44.0 | -2.0 | 0.8407 | ns |
| | | GPQA | 2000→4000 | 44.0 | 42.0 | -2.0 | 0.8399 | ns |
| | | GPQA | 300→500 | 50.0 | 40.0 | -10.0 | 0.3149 | ns |
| | | GPQA | 500→1000 | 40.0 | 46.0 | +6.0 | 0.5445 | ns |
| | | GSM8K | 1000→2000 | 38.0 | 46.0 | +8.0 | 0.4177 | ns |
| | | GSM8K | 2000→4000 | 46.0 | 44.0 | -2.0 | 0.8407 | ns |
| | | GSM8K | 300→500 | 16.0 | 30.0 | +14.0 | 0.0962 | ns |
| | | GSM8K | 500→1000 | 30.0 | 38.0 | +8.0 | 0.3984 | ns |
| DS-R1-14B | Capacity Ltd | ARC-Easy | 1000→2000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 70.0 | 90.0 | +20.0 | 0.0124 | * |
| | | ARC-Easy | 500→1000 | 90.0 | 100.0 | +10.0 | 0.0218 | * |
| | | BoolQ | 1000→2000 | 84.0 | 84.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 84.0 | 84.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 80.0 | 82.0 | +2.0 | 0.7988 | ns |
| | | BoolQ | 500→1000 | 82.0 | 84.0 | +2.0 | 0.7901 | ns |
| | | GPQA | 1000→2000 | 64.0 | 72.0 | +8.0 | 0.3912 | ns |
| | | GPQA | 2000→4000 | 72.0 | 74.0 | +2.0 | 0.8218 | ns |
| | | GPQA | 300→500 | 46.0 | 50.0 | +4.0 | 0.6889 | ns |
| | | GPQA | 500→1000 | 50.0 | 64.0 | +14.0 | 0.1574 | ns |
| | | GSM8K | 1000→2000 | 70.0 | 82.0 | +12.0 | 0.1601 | ns |
| | | GSM8K | 2000→4000 | 82.0 | 78.0 | -4.0 | 0.6171 | ns |
| | | GSM8K | 300→500 | 42.0 | 48.0 | +6.0 | 0.5465 | ns |
| | | GSM8K | 500→1000 | 48.0 | 70.0 | +22.0 | 0.0253 | * |
| DS-R1-7B | Capacity Ltd | ARC-Easy | 1000→2000 | 92.0 | 94.0 | +2.0 | 0.6951 | ns |
| | | ARC-Easy | 2000→4000 | 94.0 | 96.0 | +2.0 | 0.6464 | ns |

Table 15 (continued)

| Model | Profile | Dataset | Comparison | Before (%) | After (%) | Δ (pp) | $p$-value | Sig. |
|---|---|---|---|---|---|---|---|---|
| | | ARC-Easy | 300→500 | 34.0 | 60.0 | +26.0 | 0.0092 | ** |
| | | ARC-Easy | 500→1000 | 60.0 | 92.0 | +32.0 | 0.0002 | *** |
| | | BoolQ | 1000→2000 | 68.0 | 68.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 68.0 | 68.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 64.0 | 68.0 | +4.0 | 0.6729 | ns |
| | | BoolQ | 500→1000 | 68.0 | 68.0 | +0.0 | 1.0000 | ns |
| | | GPQA | 1000→2000 | 70.0 | 68.0 | -2.0 | 0.8288 | ns |
| | | GPQA | 2000→4000 | 68.0 | 66.0 | -2.0 | 0.8316 | ns |
| | | GPQA | 300→500 | 62.0 | 54.0 | -8.0 | 0.4177 | ns |
| | | GPQA | 500→1000 | 54.0 | 70.0 | +16.0 | 0.0993 | ns |
| | | GSM8K | 1000→2000 | 52.0 | 66.0 | +14.0 | 0.1547 | ns |
| | | GSM8K | 2000→4000 | 66.0 | 66.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 300→500 | 12.0 | 24.0 | +12.0 | 0.1183 | ns |
| | | GSM8K | 500→1000 | 24.0 | 52.0 | +28.0 | 0.0039 | ** |
| Kimi2.5-1T | Capacity Ltd | ARC-Easy | 1000→2000 | 92.0 | 100.0 | +8.0 | 0.0412 | * |
| | | ARC-Easy | 2000→4000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 24.0 | 28.0 | +4.0 | 0.6484 | ns |
| | | ARC-Easy | 4000→8000 | 100.0 | 100.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 500→1000 | 28.0 | 92.0 | +64.0 | 0.0000 | *** |
| | | BoolQ | 1000→2000 | 2.0 | 2.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 2000→4000 | 2.0 | 0.0 | -2.0 | 0.3149 | ns |
| | | BoolQ | 300→500 | 12.0 | 4.0 | -8.0 | 0.1404 | ns |
| | | BoolQ | 4000→8000 | 0.0 | 0.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 4.0 | 2.0 | -2.0 | 0.5577 | ns |
| | | GPQA | 1000→2000 | 22.0 | 28.0 | +6.0 | 0.4884 | ns |
| | | GPQA | 2000→4000 | 28.0 | 40.0 | +12.0 | 0.2053 | ns |
| | | GPQA | 300→500 | 16.0 | 26.0 | +10.0 | 0.2196 | ns |
| | | GPQA | 4000→8000 | 40.0 | 60.0 | +20.0 | 0.0455 | * |
| | | GPQA | 500→1000 | 26.0 | 22.0 | -4.0 | 0.6396 | ns |
| | | GSM8K | 1000→2000 | 60.0 | 62.0 | +2.0 | 0.8376 | ns |
| | | GSM8K | 2000→4000 | 62.0 | 72.0 | +10.0 | 0.2876 | ns |
| | | GSM8K | 300→500 | 18.0 | 48.0 | +30.0 | 0.0014 | ** |
| | | GSM8K | 4000→8000 | 72.0 | 68.0 | -4.0 | 0.6625 | ns |
| | | GSM8K | 500→1000 | 48.0 | 60.0 | +12.0 | 0.2286 | ns |
| Marco-o1-7B | Capacity Ltd | ARC-Easy | 1000→2000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 2000→4000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | ARC-Easy | 300→500 | 40.0 | 96.0 | +56.0 | 0.0000 | *** |
| | | ARC-Easy | 500→1000 | 96.0 | 96.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 1000→2000 | 90.0 | 92.0 | +2.0 | 0.7268 | ns |
| | | BoolQ | 2000→4000 | 92.0 | 92.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 300→500 | 92.0 | 92.0 | +0.0 | 1.0000 | ns |
| | | BoolQ | 500→1000 | 92.0 | 90.0 | -2.0 | 0.7268 | ns |
| | | GPQA | 1000→2000 | 30.0 | 26.0 | -4.0 | 0.6560 | ns |
| | | GPQA | 2000→4000 | 26.0 | 28.0 | +2.0 | 0.8218 | ns |
| | | GPQA | 300→500 | 44.0 | 54.0 | +10.0 | 0.3172 | ns |
| | | GPQA | 500→1000 | 54.0 | 30.0 | -24.0 | 0.0150 | * |
| | | GSM8K | 1000→2000 | 56.0 | 54.0 | -2.0 | 0.8407 | ns |
| | | GSM8K | 2000→4000 | 54.0 | 54.0 | +0.0 | 1.0000 | ns |
| | | GSM8K | 300→500 | 16.0 | 34.0 | +18.0 | 0.0377 | * |
| | | GSM8K | 500→1000 | 34.0 | 56.0 | +22.0 | 0.0270 | * |

## E.6 Token Utilisation at the 300-Token Limit

Table S16 reports, for the most restrictive token condition (300 tokens), the proportion of responses that reached the hard generation ceiling ("hit limit") and the mean number of tokens generated. A hit-limit percentage of 100% indicates that the model consistently exhausted the budget, strongly implying that the observed accuracy reflects a truncated rather than complete reasoning chain.

All 13 models exhibit severe constraint (hit-limit $\geq$ 90%) on at least one dataset at 300 tokens. The DeepSeek-R1-685B results are particularly notable: the model generates an average of 6 455 tokens on GPQA under unconstrained conditions and does not comply with the 300-token limit, so its 44% accuracy at this nominal limit is not a constrained result (compliant peak: 40% at 8 000 tokens; Section D.10.5). These results confirm that 300-token performance is not a reliable measure of model capability and justify the use of extended token budgets in the main analysis.

Table 16: Token utilisation statistics under the most restrictive condition (300-token limit). "Hit Limit" = percentage of responses that reached the token ceiling. "Avg Tokens Used" = mean generation length across 50 samples.

| Model | Profile | Dataset | Accuracy (%) | Hit Limit (%) | Avg Tokens |
|---|---|---|---|---|---|
| DeepSeek-R1-685B | Late Bloomer | ARC-Easy | 78.0 | 100.0 | 1003.2‡ |
| | | BoolQ | 30.0 | 72.0 | 277.4 |
| | | GPQA | 44.0 | 100.0 | 6455.5‡ |
| | | GSM8K | 24.0 | 100.0 | 300.0 |
| GLM-4.7-355B | Late Bloomer | ARC-Easy | 28.0 | 100.0 | 300.0 |
| | | BoolQ | 20.0 | 98.0 | 299.6 |
| | | GPQA | 14.0 | 100.0 | 300.0 |
| | | GSM8K | 8.0 | 100.0 | 300.0 |
| Nemotron-7B | Late Bloomer | ARC-Easy | 74.0 | 90.0 | 293.7 |
| | | BoolQ | 62.0 | 90.0 | 292.3 |
| | | GPQA | 32.0 | 100.0 | 300.0 |
| | | GSM8K | 38.0 | 94.0 | 294.2 |
| Qwen3-32B-O1 | Late Bloomer | ARC-Easy | 42.0 | 98.0 | 298.9 |
| | | BoolQ | 32.0 | 24.0 | 203.6 |
| | | GPQA | 24.0 | 100.0 | 300.0 |
| | | GSM8K | 16.0 | 100.0 | 300.0 |
| DS-R1-70B | Early Plateau | ARC-Easy | 26.0 | 98.0 | 298.4 |
| | | BoolQ | 88.0 | 60.0 | 268.4 |
| | | GPQA | 54.0 | 100.0 | 300.0 |
| | | GSM8K | 56.0 | 98.0 | 299.2 |
| QwenQwQ-32B | Early Plateau | ARC-Easy | 62.0 | 48.0 | 242.8 |
| | | BoolQ | 88.0 | 6.0 | 69.9 |
| | | GPQA | 34.0 | 96.0 | 295.8 |
| | | GSM8K | 34.0 | 82.0 | 284.3 |
| Skywork-o1-8B | Early Plateau | ARC-Easy | 56.0 | 54.0 | 250.1 |
| | | BoolQ | 54.0 | 12.0 | 102.6 |
| | | GPQA | 40.0 | 96.0 | 297.3 |
| | | GSM8K | 58.0 | 48.0 | 258.4 |
| GLM-4.5 | Task-Specific | ARC-Easy | 30.0 | 100.0 | 300.0 |
| | | BoolQ | 12.0 | 100.0 | 299.7 |
| | | GPQA | 20.0 | 100.0 | 300.0 |
| | | GSM8K | 4.0 | 100.0 | 300.0 |
| DS-R1-1.5B | Capacity Ltd | ARC-Easy | 38.0 | 40.0 | 236.4 |
| | | BoolQ | 38.0 | 38.0 | 160.7 |
| | | GPQA | 50.0 | 98.0 | 296.9 |
| | | GSM8K | 16.0 | 100.0 | 300.0 |
| DS-R1-14B | Capacity Ltd | ARC-Easy | 70.0 | 46.0 | 233.0 |
| | | BoolQ | 80.0 | 6.0 | 44.4 |
| | | GPQA | 46.0 | 94.0 | 295.3 |
| | | GSM8K | 42.0 | 90.0 | 296.0 |
| DS-R1-7B | Capacity Ltd | ARC-Easy | 34.0 | 86.0 | 289.6 |
| | | BoolQ | 64.0 | 62.0 | 260.9 |
| | | GPQA | 62.0 | 100.0 | 300.0 |
| | | GSM8K | 12.0 | 100.0 | 300.0 |
| Kimi2.5-1T | Capacity Ltd | ARC-Easy | 24.0 | 100.0 | 300.0 |
| | | BoolQ | 12.0 | 42.0 | 229.3 |
| | | GPQA | 16.0 | 100.0 | 300.0 |
| | | GSM8K | 18.0 | 94.0 | 296.6 |
| Marco-o1-7B | Capacity Ltd | ARC-Easy | 40.0 | 100.0 | 300.0 |
| | | BoolQ | 92.0 | 58.0 | 198.4 |
| | | GPQA | 44.0 | 100.0 | 300.0 |
| | | GSM8K | 16.0 | 98.0 | 298.8 |

‡DeepSeek-R1-685B generates extended internal reasoning traces that exceed stated token limits on ARC-Easy and GPQA at low token budgets; "Avg Tokens" reports actual generation length. See Section D.10.5 for complete impact assessment and compliance thresholds.

## E.7 Statistical Notes

**Sample size.** Each accuracy estimate is based on $n = 50$ randomly sampled questions per condition. This sample size is consistent with established practice in rapid evaluation benchmarking but limits the detection power for small effects: the minimum detectable effect at $\alpha = 0.05$ and power $= 0.80$ for McNemar's test is approximately $\Delta \approx 18$ pp under this design. Readers should interpret improvements of $< 10$ pp with caution regardless of reported $p$-values.

**Correction for multiple comparisons.** A total of 236 pairwise tests were conducted (Table S15). Main-text claims are restricted to effects that are (a) statistically significant at $p < 0.05$ uncorrected, and (b) of practical magnitude ($\Delta \geq 20$ pp or $|h| \geq 0.40$). The primary conclusions—the +36 pp and +20 pp gains for DeepSeek-R1-685B on GSM8K ($p < 0.001$ for both), the +24 pp gain for Qwen3-32B-O1 ($p = 0.001$), and the +60 pp and +50 pp ARC-Easy gains for GLM-4.5 and GLM-4.7-355B ($p < 0.001$ for both)—remain significant under a Bonferroni-corrected threshold of $\alpha' = 0.05/236 = 0.00021$.

**GPQA degradation.** The over-generation criterion (peak at $\leq 1\,000$ tokens, net drop $\geq 6$ pp) identified three models (23% of the evaluated set). The most pronounced individual step-wise effect is the $-24$ pp decline in Marco-o1-7B GPQA accuracy from 500 to 1 000 tokens ($p = 0.015$, McNemar's test). Subsequent step-wise changes for this model do not reach statistical significance ($p > 0.65$), indicating that the primary degradation occurs at the 500-to-1 000-token boundary.

**BoolQ anomaly.** GLM-4.5, GLM-4.7-355B (active 32B), and Kimi2.5-1T (active 32B) record near-zero or declining BoolQ accuracy across all token limits (0–12%). This pattern is not consistent with a token-budget effect and is likely attributable to a mismatch between the model's output format and the binary-answer extraction heuristic. These three models are excluded from the BoolQ plateau count reported in the main text (9 of 10 eligible models).

# F Training Quality Efficiency Improvement at Matched Parameter Scales

## F.1 Overview

This note details the calculation methodology for the 18.8-fold efficiency improvement observed when comparing models with excellent versus poor training quality at matched parameter scales (7–8B). The comparison directly isolates training methodology as the dominant factor in test-time reasoning efficiency.

## F.2 Calculation Methodology

We quantify token efficiency using the metric:

$$\text{Efficiency} = \frac{\text{Peak Accuracy}}{\text{Average Tokens Consumed}} \tag{8}$$

This metric measures how effectively a model converts computational resources (tokens generated) into task performance (accuracy achieved). Higher values indicate better token efficiency—achieving more accuracy per token consumed.

## F.3 Experimental Data

Token ablation experiments on GSM8K (50 samples per condition, up to 6 token limits: 300, 500, 1000, 2000, 4000, 8000; small models (1.5B–14B) were evaluated at 300–4000 and large models (32B–685B) at 300–8000) yield the following results for models at matched 7–8B parameter scales:

**Skywork-o1-8B (Excellent Training, Tier 1):**

- Training quality: $Q = 0.987$
- Peak accuracy: 86.0% (achieved at 1000-token limit)
- Average tokens consumed at plateau: 292.84 tokens
- Token budget utilization: 0.0% (plateau reached without hitting limit)
- Efficiency: $0.86 \div 292.84 = 0.002937$ accuracy points per token

**Marco-o1-7B (Poor Training, Tier 5):**

- Training quality: $Q = 0.243$
- Peak accuracy: 56.0% (at 1000-token limit)
- Accuracy at 4000-token limit: 54.0% (performance degradation beyond plateau)
- Average tokens consumed at 4000 limit: 3464.80 tokens
- Token budget utilization: 82.0% (compensatory verbosity)
- Efficiency at 4000 tokens: $0.54 \div 3464.80 = 0.000156$ accuracy points per token

## F.4 Efficiency Improvement Calculation

The efficiency improvement ratio at matched parameter scales (7–8B) is:

$$\text{Improvement Ratio} = \frac{\text{Efficiency}_{\text{Skywork}}}{\text{Efficiency}_{\text{Marco}}} = \frac{0.002937}{0.000156} = 18.8\times \tag{9}$$

## F.5 Interpretation

This 18.8-fold improvement demonstrates three critical findings:

1. **Training dominates parameter scaling:** Skywork-o1-8B achieves higher accuracy (86% vs 54%) with dramatically fewer tokens (293 vs 3465), establishing both quality and efficiency advantages through training optimization alone.
2. **Exceeds parameter scaling gains:** The efficiency improvement from excellent training at 8B parameters exceeds the typical gains from $10\times$ parameter scaling within poor-training regimes. Marco-o1-7B performance at 7B parameters would require similar inefficiency at 70B parameters without training improvements.
3. **Persistent across token budgets:** Even at Marco's peak performance (56% at 1000 tokens using 961.90 tokens), Skywork achieves $0.002937 \div (0.56 \div 961.90) = 5.0\times$ better efficiency, confirming the advantage persists independent of token budget.

## F.6 Robustness Validation

The calculation remains consistent across alternative formulations:

- **Token usage method:** Using token budget utilization percentages yields $(0.86/0.073)/(0.54/0.866) = 18.9\times$
- **Intermediate models validate gradient:** Nemotron-7B (moderate training, $Q = 0.556$) achieves 80% accuracy at 4000 tokens (efficiency = 0.00044), placing it precisely between Skywork and Marco on the training quality gradient.
- **Task dependence:** The efficiency advantage is strongest on GSM8K; on ARC-Easy and BoolQ, Marco-o1-7B achieves higher peak accuracy despite lower Q-score, consistent with the cross-dataset Q-score shifts documented in Appendix H.3.

Table 17: **Complete Token Ablation Data for Training Quality Analysis.** Systematic evaluation of 13 reasoning-specialized models across 4 datasets and 6 token limits (300–8000 for flagship models; 300–4000 for small models). Models span 1.5B to 685B parameters with diverse training methodologies. Key comparisons: Skywork-o1-8B (excellent training, $Q = 0.987$) vs Marco-o1-7B (poor training, $Q = 0.243$) at matched 7–8B scales demonstrate $18.8\times$ efficiency improvement. Each condition evaluated on 50 samples. All 288 experimental conditions shown.

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| **DeepSeek-R1-Distill-Qwen-1.5B (1.5B)** | | | | | | |
| DR1-Qwen-1.5B | 1.5 | arc-easy | 300 | 38.0 | 236.4 | 40.0 |
| DR1-Qwen-1.5B | 1.5 | arc-easy | 500 | 42.0 | 282.8 | 20.0 |
| DR1-Qwen-1.5B | 1.5 | arc-easy | 1000 | 48.0 | 347.8 | 10.0 |
| DR1-Qwen-1.5B | 1.5 | arc-easy | 2000 | 48.0 | 447.8 | 10.0 |
| DR1-Qwen-1.5B | 1.5 | arc-easy | 4000 | 48.0 | 647.8 | 10.0 |
| DR1-Qwen-1.5B | 1.5 | boolq | 300 | 38.0 | 160.7 | 38.0 |
| DR1-Qwen-1.5B | 1.5 | boolq | 500 | 44.0 | 231.5 | 32.0 |
| DR1-Qwen-1.5B | 1.5 | boolq | 1000 | 44.0 | 391.5 | 32.0 |
| DR1-Qwen-1.5B | 1.5 | boolq | 2000 | 44.0 | 711.5 | 32.0 |
| DR1-Qwen-1.5B | 1.5 | boolq | 4000 | 44.0 | 1351.5 | 32.0 |
| DR1-Qwen-1.5B | 1.5 | gpqa | 300 | 50.0 | 296.9 | 98.0 |
| DR1-Qwen-1.5B | 1.5 | gpqa | 500 | 40.0 | 449.0 | 60.0 |
| DR1-Qwen-1.5B | 1.5 | gpqa | 1000 | 46.0 | 686.0 | 36.0 |
| DR1-Qwen-1.5B | 1.5 | gpqa | 2000 | 44.0 | 1034.2 | 34.0 |
| DR1-Qwen-1.5B | 1.5 | gpqa | 4000 | 42.0 | 1714.2 | 34.0 |
| DR1-Qwen-1.5B | 1.5 | gsm8k | 300 | 16.0 | 300.0 | 100.0 |
| DR1-Qwen-1.5B | 1.5 | gsm8k | 500 | 30.0 | 500.0 | 100.0 |
| DR1-Qwen-1.5B | 1.5 | gsm8k | 1000 | 38.0 | 999.5 | 98.0 |
| DR1-Qwen-1.5B | 1.5 | gsm8k | 2000 | 46.0 | 1957.6 | 94.0 |

*Continued on next page*

*Table 17 continued from previous page*

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| DR1-Qwen-1.5B | 1.5 | gsm8k | 4000 | 44.0 | 3837.6 | 94.0 |
| **Marco-o1-7B (7B) - Tier 5: Insufficient Training** | | | | | | |
| Marco-o1-7B | 7.0 | arc-easy | 300 | 40.0 | 300.0 | 100.0 |
| Marco-o1-7B | 7.0 | arc-easy | 500 | 96.0 | 496.6 | 98.0 |
| Marco-o1-7B | 7.0 | arc-easy | 1000 | 96.0 | 986.6 | 98.0 |
| Marco-o1-7B | 7.0 | arc-easy | 2000 | 96.0 | 1966.6 | 98.0 |
| Marco-o1-7B | 7.0 | arc-easy | 4000 | 96.0 | 3926.6 | 98.0 |
| Marco-o1-7B | 7.0 | boolq | 300 | 92.0 | 198.4 | 58.0 |
| Marco-o1-7B | 7.0 | boolq | 500 | 92.0 | 314.4 | 58.0 |
| Marco-o1-7B | 7.0 | boolq | 1000 | 90.0 | 596.0 | 56.0 |
| Marco-o1-7B | 7.0 | boolq | 2000 | 92.0 | 1223.2 | 60.0 |
| Marco-o1-7B | 7.0 | boolq | 4000 | 92.0 | 2423.2 | 60.0 |
| Marco-o1-7B | 7.0 | gpqa | 300 | 44.0 | 300.0 | 100.0 |
| Marco-o1-7B | 7.0 | gpqa | 500 | 54.0 | 499.3 | 98.0 |
| Marco-o1-7B | 7.0 | gpqa | 1000 | 30.0 | 989.3 | 98.0 |
| Marco-o1-7B | 7.0 | gpqa | 2000 | 26.0 | 1969.3 | 98.0 |
| Marco-o1-7B | 7.0 | gpqa | 4000 | 28.0 | 3929.3 | 98.0 |
| Marco-o1-7B | 7.0 | gsm8k | 300 | 16.0 | 298.8 | 98.0 |
| Marco-o1-7B | 7.0 | gsm8k | 500 | 34.0 | 494.8 | 98.0 |
| Marco-o1-7B | 7.0 | gsm8k | 1000 | 56.0 | 961.9 | 92.0 |
| Marco-o1-7B | 7.0 | gsm8k | 2000 | 54.0 | 1853.8 | 88.0 |
| Marco-o1-7B | 7.0 | gsm8k | 4000 | 54.0 | 3464.8 | 82.0 |
| **DeepSeek-R1-Distill-Qwen-7B (7B)** | | | | | | |
| DR1-Qwen-7B | 7.0 | arc-easy | 300 | 34.0 | 289.6 | 86.0 |
| DR1-Qwen-7B | 7.0 | arc-easy | 500 | 60.0 | 429.9 | 52.0 |
| DR1-Qwen-7B | 7.0 | arc-easy | 1000 | 92.0 | 570.8 | 18.0 |
| DR1-Qwen-7B | 7.0 | arc-easy | 2000 | 94.0 | 681.3 | 10.0 |
| DR1-Qwen-7B | 7.0 | arc-easy | 4000 | 96.0 | 786.4 | 4.0 |
| DR1-Qwen-7B | 7.0 | boolq | 300 | 64.0 | 260.9 | 62.0 |
| DR1-Qwen-7B | 7.0 | boolq | 500 | 68.0 | 364.4 | 46.0 |
| DR1-Qwen-7B | 7.0 | boolq | 1000 | 68.0 | 532.8 | 28.0 |
| DR1-Qwen-7B | 7.0 | boolq | 2000 | 68.0 | 798.2 | 26.0 |
| DR1-Qwen-7B | 7.0 | boolq | 4000 | 68.0 | 1318.2 | 26.0 |
| DR1-Qwen-7B | 7.0 | gpqa | 300 | 62.0 | 300.0 | 100.0 |
| DR1-Qwen-7B | 7.0 | gpqa | 500 | 54.0 | 499.1 | 98.0 |
| DR1-Qwen-7B | 7.0 | gpqa | 1000 | 70.0 | 975.2 | 90.0 |
| DR1-Qwen-7B | 7.0 | gpqa | 2000 | 68.0 | 1750.2 | 72.0 |
| DR1-Qwen-7B | 7.0 | gpqa | 4000 | 66.0 | 3206.7 | 66.0 |
| DR1-Qwen-7B | 7.0 | gsm8k | 300 | 12.0 | 300.0 | 100.0 |
| DR1-Qwen-7B | 7.0 | gsm8k | 500 | 24.0 | 500.0 | 100.0 |
| DR1-Qwen-7B | 7.0 | gsm8k | 1000 | 52.0 | 958.6 | 86.0 |
| DR1-Qwen-7B | 7.0 | gsm8k | 2000 | 66.0 | 1615.2 | 56.0 |
| DR1-Qwen-7B | 7.0 | gsm8k | 4000 | 66.0 | 2641.6 | 50.0 |
| **Nemotron-7B (7B) - Tier 4: Poor Training** | | | | | | |
| Nemotron-7B | 7.0 | arc-easy | 300 | 74.0 | 293.7 | 90.0 |
| Nemotron-7B | 7.0 | arc-easy | 500 | 78.0 | 435.2 | 44.0 |
| Nemotron-7B | 7.0 | arc-easy | 1000 | 78.0 | 568.4 | 20.0 |
| Nemotron-7B | 7.0 | arc-easy | 2000 | 84.0 | 673.0 | 2.0 |
| Nemotron-7B | 7.0 | arc-easy | 4000 | 86.0 | 713.0 | 2.0 |
| Nemotron-7B | 7.0 | boolq | 300 | 62.0 | 292.3 | 90.0 |
| Nemotron-7B | 7.0 | boolq | 500 | 68.0 | 447.9 | 74.0 |
| Nemotron-7B | 7.0 | boolq | 1000 | 68.0 | 765.8 | 48.0 |
| Nemotron-7B | 7.0 | boolq | 2000 | 70.0 | 1185.7 | 38.0 |
| Nemotron-7B | 7.0 | boolq | 4000 | 70.0 | 1762.0 | 26.0 |
| Nemotron-7B | 7.0 | gpqa | 300 | 32.0 | 300.0 | 100.0 |
| Nemotron-7B | 7.0 | gpqa | 500 | 44.0 | 499.9 | 100.0 |
| Nemotron-7B | 7.0 | gpqa | 1000 | 58.0 | 972.6 | 92.0 |
| Nemotron-7B | 7.0 | gpqa | 2000 | 46.0 | 1861.8 | 86.0 |
| Nemotron-7B | 7.0 | gpqa | 4000 | 52.0 | 3408.7 | 68.0 |
| Nemotron-7B | 7.0 | gsm8k | 300 | 38.0 | 294.2 | 94.0 |

*Continued on next page*

*Table 17 continued from previous page*

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| Nemotron-7B | 7.0 | gsm8k | 500 | 42.0 | 465.4 | 78.0 |
| Nemotron-7B | 7.0 | gsm8k | 1000 | 52.0 | 794.8 | 56.0 |
| Nemotron-7B | 7.0 | gsm8k | 2000 | 64.0 | 1296.0 | 44.0 |
| Nemotron-7B | 7.0 | gsm8k | 4000 | 80.0 | 1816.9 | 16.0 |
| **Skywork-o1-8B (8B) - Tier 1: Excellent Training** | | | | | | |
| Skywork-o1-8B | 8.0 | arc-easy | 300 | 56.0 | 250.1 | 54.0 |
| Skywork-o1-8B | 8.0 | arc-easy | 500 | 60.0 | 342.8 | 44.0 |
| Skywork-o1-8B | 8.0 | arc-easy | 1000 | 60.0 | 553.3 | 42.0 |
| Skywork-o1-8B | 8.0 | arc-easy | 2000 | 60.0 | 973.3 | 42.0 |
| Skywork-o1-8B | 8.0 | arc-easy | 4000 | 60.0 | 1813.3 | 42.0 |
| Skywork-o1-8B | 8.0 | boolq | 300 | 54.0 | 102.6 | 12.0 |
| Skywork-o1-8B | 8.0 | boolq | 500 | 54.0 | 126.6 | 12.0 |
| Skywork-o1-8B | 8.0 | boolq | 1000 | 54.0 | 186.6 | 12.0 |
| Skywork-o1-8B | 8.0 | boolq | 2000 | 54.0 | 306.6 | 12.0 |
| Skywork-o1-8B | 8.0 | boolq | 4000 | 54.0 | 546.6 | 12.0 |
| Skywork-o1-8B | 8.0 | gpqa | 300 | 40.0 | 297.3 | 96.0 |
| Skywork-o1-8B | 8.0 | gpqa | 500 | 42.0 | 464.1 | 72.0 |
| Skywork-o1-8B | 8.0 | gpqa | 1000 | 38.0 | 736.2 | 46.0 |
| Skywork-o1-8B | 8.0 | gpqa | 2000 | 38.0 | 1074.0 | 32.0 |
| Skywork-o1-8B | 8.0 | gpqa | 4000 | 36.0 | 1714.0 | 32.0 |
| Skywork-o1-8B | 8.0 | gsm8k | 300 | 58.0 | 258.4 | 48.0 |
| Skywork-o1-8B | 8.0 | gsm8k | 500 | 82.0 | 290.9 | 4.0 |
| Skywork-o1-8B | 8.0 | gsm8k | 1000 | 86.0 | 292.8 | 0.0 |
| Skywork-o1-8B | 8.0 | gsm8k | 2000 | 86.0 | 292.8 | 0.0 |
| Skywork-o1-8B | 8.0 | gsm8k | 4000 | 86.0 | 292.8 | 0.0 |
| **GLM-4.5 (355B total, 32B active) - Tier 4: Poor Training** | | | | | | |
| GLM-4.5 | 32.0 | arc-easy | 300 | 30.0 | 300.0 | 100.0 |
| GLM-4.5 | 32.0 | arc-easy | 500 | 12.0 | 500.0 | 100.0 |
| GLM-4.5 | 32.0 | arc-easy | 1000 | 40.0 | 969.4 | 74.0 |
| GLM-4.5 | 32.0 | arc-easy | 2000 | 100.0 | 1184.6 | 0.0 |
| GLM-4.5 | 32.0 | arc-easy | 4000 | 98.0 | 1194.2 | 0.0 |
| GLM-4.5 | 32.0 | arc-easy | 8000 | 100.0 | 1165.5 | 0.0 |
| GLM-4.5 | 32.0 | boolq | 300 | 12.0 | 299.7 | 100.0 |
| GLM-4.5 | 32.0 | boolq | 500 | 10.0 | 425.9 | 42.0 |
| GLM-4.5 | 32.0 | boolq | 1000 | 8.0 | 572.4 | 20.0 |
| GLM-4.5 | 32.0 | boolq | 2000 | 6.0 | 742.7 | 14.0 |
| GLM-4.5 | 32.0 | boolq | 4000 | 0.0 | 863.8 | 4.0 |
| GLM-4.5 | 32.0 | boolq | 8000 | 0.0 | 914.8 | 0.0 |
| GLM-4.5 | 32.0 | gpqa | 300 | 20.0 | 300.0 | 100.0 |
| GLM-4.5 | 32.0 | gpqa | 500 | 28.0 | 500.0 | 100.0 |
| GLM-4.5 | 32.0 | gpqa | 1000 | 22.0 | 1000.0 | 100.0 |
| GLM-4.5 | 32.0 | gpqa | 2000 | 26.0 | 1973.9 | 94.0 |
| GLM-4.5 | 32.0 | gpqa | 4000 | 32.0 | 3724.5 | 74.0 |
| GLM-4.5 | 32.0 | gpqa | 8000 | 66.0 | 5852.6 | 34.0 |
| GLM-4.5 | 32.0 | gsm8k | 300 | 4.0 | 300.0 | 100.0 |
| GLM-4.5 | 32.0 | gsm8k | 500 | 20.0 | 493.3 | 94.0 |
| GLM-4.5 | 32.0 | gsm8k | 1000 | 66.0 | 815.5 | 36.0 |
| GLM-4.5 | 32.0 | gsm8k | 2000 | 74.0 | 962.9 | 6.0 |
| GLM-4.5 | 32.0 | gsm8k | 4000 | 78.0 | 1041.7 | 2.0 |
| GLM-4.5 | 32.0 | gsm8k | 8000 | 80.0 | 965.3 | 0.0 |
| **DeepSeek-R1-Distill-Qwen-14B (14B)** | | | | | | |
| DR1-Qwen-14B | 14.0 | arc-easy | 300 | 70.0 | 233.0 | 46.0 |
| DR1-Qwen-14B | 14.0 | arc-easy | 500 | 90.0 | 281.9 | 14.0 |
| DR1-Qwen-14B | 14.0 | arc-easy | 1000 | 100.0 | 308.2 | 2.0 |
| DR1-Qwen-14B | 14.0 | arc-easy | 2000 | 100.0 | 310.5 | 0.0 |
| DR1-Qwen-14B | 14.0 | arc-easy | 4000 | 100.0 | 310.5 | 0.0 |
| DR1-Qwen-14B | 14.0 | boolq | 300 | 80.0 | 44.4 | 6.0 |
| DR1-Qwen-14B | 14.0 | boolq | 500 | 82.0 | 49.7 | 2.0 |
| DR1-Qwen-14B | 14.0 | boolq | 1000 | 84.0 | 59.7 | 2.0 |
| DR1-Qwen-14B | 14.0 | boolq | 2000 | 84.0 | 75.9 | 0.0 |

*Continued on next page*

*Table 17 continued from previous page*

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| DR1-Qwen-14B | 14.0 | boolq | 4000 | 84.0 | 75.9 | 0.0 |
| DR1-Qwen-14B | 14.0 | gpqa | 300 | 46.0 | 295.3 | 94.0 |
| DR1-Qwen-14B | 14.0 | gpqa | 500 | 50.0 | 475.1 | 84.0 |
| DR1-Qwen-14B | 14.0 | gpqa | 1000 | 64.0 | 845.0 | 66.0 |
| DR1-Qwen-14B | 14.0 | gpqa | 2000 | 72.0 | 1425.3 | 50.0 |
| DR1-Qwen-14B | 14.0 | gpqa | 4000 | 74.0 | 2280.4 | 36.0 |
| DR1-Qwen-14B | 14.0 | gsm8k | 300 | 42.0 | 296.0 | 90.0 |
| DR1-Qwen-14B | 14.0 | gsm8k | 500 | 48.0 | 459.1 | 74.0 |
| DR1-Qwen-14B | 14.0 | gsm8k | 1000 | 70.0 | 723.6 | 32.0 |
| DR1-Qwen-14B | 14.0 | gsm8k | 2000 | 82.0 | 914.9 | 12.0 |
| DR1-Qwen-14B | 14.0 | gsm8k | 4000 | 78.0 | 1238.6 | 16.0 |
| **QwQ-32B-Preview (32B) - Tier 1: Excellent Training** | | | | | | |
| QwQ-32B | 32.0 | arc-easy | 300 | 62.0 | 242.8 | 48.0 |
| QwQ-32B | 32.0 | arc-easy | 500 | 86.0 | 285.1 | 8.0 |
| QwQ-32B | 32.0 | arc-easy | 1000 | 90.0 | 302.0 | 2.0 |
| QwQ-32B | 32.0 | arc-easy | 2000 | 90.0 | 310.4 | 0.0 |
| QwQ-32B | 32.0 | arc-easy | 4000 | 90.0 | 310.4 | 0.0 |
| QwQ-32B | 32.0 | arc-easy | 8000 | 90.0 | 310.4 | 0.0 |
| QwQ-32B | 32.0 | boolq | 300 | 88.0 | 69.9 | 6.0 |
| QwQ-32B | 32.0 | boolq | 500 | 88.0 | 75.5 | 2.0 |
| QwQ-32B | 32.0 | boolq | 1000 | 88.0 | 85.5 | 2.0 |
| QwQ-32B | 32.0 | boolq | 2000 | 88.0 | 105.5 | 2.0 |
| QwQ-32B | 32.0 | boolq | 4000 | 88.0 | 145.5 | 2.0 |
| QwQ-32B | 32.0 | boolq | 8000 | 88.0 | 225.5 | 2.0 |
| QwQ-32B | 32.0 | gpqa | 300 | 34.0 | 295.8 | 96.0 |
| QwQ-32B | 32.0 | gpqa | 500 | 56.0 | 481.1 | 86.0 |
| QwQ-32B | 32.0 | gpqa | 1000 | 54.0 | 810.5 | 48.0 |
| QwQ-32B | 32.0 | gpqa | 2000 | 58.0 | 1173.0 | 26.0 |
| QwQ-32B | 32.0 | gpqa | 4000 | 64.0 | 1486.7 | 6.0 |
| QwQ-32B | 32.0 | gpqa | 8000 | 64.0 | 1587.5 | 2.0 |
| QwQ-32B | 32.0 | gsm8k | 300 | 34.0 | 284.3 | 82.0 |
| QwQ-32B | 32.0 | gsm8k | 500 | 82.0 | 394.2 | 30.0 |
| QwQ-32B | 32.0 | gsm8k | 1000 | 92.0 | 447.5 | 6.0 |
| QwQ-32B | 32.0 | gsm8k | 2000 | 98.0 | 493.6 | 0.0 |
| QwQ-32B | 32.0 | gsm8k | 4000 | 98.0 | 493.6 | 0.0 |
| QwQ-32B | 32.0 | gsm8k | 8000 | 98.0 | 493.6 | 0.0 |
| **GLM-4.7 (32B, 355B total, 32B active) - Tier 4: Poor Training** | | | | | | |
| GLM-4.7 | 32.0 | arc-easy | 300 | 28.0 | 300.0 | 100.0 |
| GLM-4.7 | 32.0 | arc-easy | 500 | 14.0 | 500.0 | 100.0 |
| GLM-4.7 | 32.0 | arc-easy | 1000 | 46.0 | 955.8 | 68.0 |
| GLM-4.7 | 32.0 | arc-easy | 2000 | 96.0 | 1134.3 | 2.0 |
| GLM-4.7 | 32.0 | arc-easy | 4000 | 96.0 | 1172.5 | 2.0 |
| GLM-4.7 | 32.0 | arc-easy | 8000 | 98.0 | 1217.3 | 0.0 |
| GLM-4.7 | 32.0 | boolq | 300 | 20.0 | 299.6 | 98.0 |
| GLM-4.7 | 32.0 | boolq | 500 | 12.0 | 416.6 | 34.0 |
| GLM-4.7 | 32.0 | boolq | 1000 | 4.0 | 528.9 | 16.0 |
| GLM-4.7 | 32.0 | boolq | 2000 | 8.0 | 652.5 | 10.0 |
| GLM-4.7 | 32.0 | boolq | 4000 | 0.0 | 778.4 | 0.0 |
| GLM-4.7 | 32.0 | boolq | 8000 | 0.0 | 696.8 | 0.0 |
| GLM-4.7 | 32.0 | gpqa | 300 | 14.0 | 300.0 | 100.0 |
| GLM-4.7 | 32.0 | gpqa | 500 | 30.0 | 500.0 | 100.0 |
| GLM-4.7 | 32.0 | gpqa | 1000 | 24.0 | 1000.0 | 100.0 |
| GLM-4.7 | 32.0 | gpqa | 2000 | 22.0 | 1980.2 | 94.0 |
| GLM-4.7 | 32.0 | gpqa | 4000 | 34.0 | 3572.5 | 64.0 |
| GLM-4.7 | 32.0 | gpqa | 8000 | 46.0 | 5406.6 | 40.0 |
| GLM-4.7 | 32.0 | gsm8k | 300 | 8.0 | 300.0 | 100.0 |
| GLM-4.7 | 32.0 | gsm8k | 500 | 8.0 | 496.2 | 96.0 |
| GLM-4.7 | 32.0 | gsm8k | 1000 | 62.0 | 819.1 | 30.0 |
| GLM-4.7 | 32.0 | gsm8k | 2000 | 76.0 | 951.7 | 6.0 |
| GLM-4.7 | 32.0 | gsm8k | 4000 | 84.0 | 937.2 | 0.0 |

*Continued on next page*

*Table 17 continued from previous page*

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| GLM-4.7 | 32.0 | gsm8k | 8000 | 86.0 | 1042.0 | 0.0 |
| **Kimi-2.5 (32B, 1T total, 32B active) - Tier 4: Poor Training** | | | | | | |
| Kimi-2.5 | 32.0 | arc-easy | 300 | 24.0 | 300.0 | 100.0 |
| Kimi-2.5 | 32.0 | arc-easy | 500 | 28.0 | 498.8 | 98.0 |
| Kimi-2.5 | 32.0 | arc-easy | 1000 | 92.0 | 747.3 | 14.0 |
| Kimi-2.5 | 32.0 | arc-easy | 2000 | 100.0 | 795.7 | 0.0 |
| Kimi-2.5 | 32.0 | arc-easy | 4000 | 100.0 | 791.6 | 0.0 |
| Kimi-2.5 | 32.0 | arc-easy | 8000 | 100.0 | 791.8 | 0.0 |
| Kimi-2.5 | 32.0 | boolq | 300 | 12.0 | 229.3 | 42.0 |
| Kimi-2.5 | 32.0 | boolq | 500 | 4.0 | 290.6 | 28.0 |
| Kimi-2.5 | 32.0 | boolq | 1000 | 2.0 | 350.4 | 10.0 |
| Kimi-2.5 | 32.0 | boolq | 2000 | 2.0 | 445.5 | 2.0 |
| Kimi-2.5 | 32.0 | boolq | 4000 | 0.0 | 478.0 | 0.0 |
| Kimi-2.5 | 32.0 | boolq | 8000 | 0.0 | 465.9 | 0.0 |
| Kimi-2.5 | 32.0 | gpqa | 300 | 16.0 | 300.0 | 100.0 |
| Kimi-2.5 | 32.0 | gpqa | 500 | 26.0 | 500.0 | 100.0 |
| Kimi-2.5 | 32.0 | gpqa | 1000 | 22.0 | 1000.0 | 100.0 |
| Kimi-2.5 | 32.0 | gpqa | 2000 | 28.0 | 1956.1 | 94.0 |
| Kimi-2.5 | 32.0 | gpqa | 4000 | 40.0 | 3608.5 | 74.0 |
| Kimi-2.5 | 32.0 | gpqa | 8000 | 60.0 | 5674.3 | 36.0 |
| Kimi-2.5 | 32.0 | gsm8k | 300 | 18.0 | 296.6 | 94.0 |
| Kimi-2.5 | 32.0 | gsm8k | 500 | 48.0 | 426.9 | 44.0 |
| Kimi-2.5 | 32.0 | gsm8k | 1000 | 60.0 | 519.9 | 8.0 |
| Kimi-2.5 | 32.0 | gsm8k | 2000 | 62.0 | 611.7 | 4.0 |
| Kimi-2.5 | 32.0 | gsm8k | 4000 | 72.0 | 597.8 | 0.0 |
| Kimi-2.5 | 32.0 | gsm8k | 8000 | 68.0 | 595.2 | 0.0 |
| **Qwen3-32B-o1 (32B) - Tier 3: Moderate Training** | | | | | | |
| Qwen3-32B-o1 | 32.0 | arc-easy | 300 | 42.0 | 298.9 | 98.0 |
| Qwen3-32B-o1 | 32.0 | arc-easy | 500 | 80.0 | 455.5 | 64.0 |
| Qwen3-32B-o1 | 32.0 | arc-easy | 1000 | 98.0 | 559.0 | 2.0 |
| Qwen3-32B-o1 | 32.0 | arc-easy | 2000 | 100.0 | 562.9 | 0.0 |
| Qwen3-32B-o1 | 32.0 | arc-easy | 4000 | 96.0 | 582.1 | 0.0 |
| Qwen3-32B-o1 | 32.0 | arc-easy | 8000 | 96.0 | 584.6 | 0.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 300 | 32.0 | 203.6 | 24.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 500 | 50.0 | 239.8 | 10.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 1000 | 44.0 | 245.4 | 0.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 2000 | 48.0 | 246.8 | 0.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 4000 | 52.0 | 250.4 | 0.0 |
| Qwen3-32B-o1 | 32.0 | boolq | 8000 | 48.0 | 224.9 | 0.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 300 | 24.0 | 300.0 | 100.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 500 | 28.0 | 500.0 | 100.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 1000 | 30.0 | 988.1 | 94.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 2000 | 44.0 | 1869.7 | 80.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 4000 | 52.0 | 3051.3 | 46.0 |
| Qwen3-32B-o1 | 32.0 | gpqa | 8000 | 44.0 | 3632.5 | 18.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 300 | 16.0 | 300.0 | 100.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 500 | 28.0 | 500.0 | 100.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 1000 | 72.0 | 960.5 | 82.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 2000 | 96.0 | 1380.0 | 14.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 4000 | 98.0 | 1583.7 | 2.0 |
| Qwen3-32B-o1 | 32.0 | gsm8k | 8000 | 98.0 | 1627.2 | 2.0 |
| **DeepSeek-R1-Distill-Llama-70B (70B) - Tier 2: Good Training** | | | | | | |
| DR1-Llama-70B | 70.0 | arc-easy | 300 | 26.0 | 298.4 | 98.0 |
| DR1-Llama-70B | 70.0 | arc-easy | 500 | 56.0 | 480.8 | 80.0 |
| DR1-Llama-70B | 70.0 | arc-easy | 1000 | 98.0 | 611.6 | 4.0 |
| DR1-Llama-70B | 70.0 | arc-easy | 2000 | 100.0 | 621.2 | 0.0 |
| DR1-Llama-70B | 70.0 | arc-easy | 4000 | 100.0 | 621.2 | 0.0 |
| DR1-Llama-70B | 70.0 | arc-easy | 8000 | 100.0 | 621.2 | 0.0 |
| DR1-Llama-70B | 70.0 | boolq | 300 | 88.0 | 268.4 | 60.0 |
| DR1-Llama-70B | 70.0 | boolq | 500 | 88.0 | 335.6 | 24.0 |

*Continued on next page*

*Table 17 continued from previous page*

| Model | Params (B) | Dataset | Limit | Acc (%) | Avg Tokens | Hit (%) |
|---|---|---|---|---|---|---|
| DR1-Llama-70B | 70.0 | boolq | 1000 | 88.0 | 411.7 | 14.0 |
| DR1-Llama-70B | 70.0 | boolq | 2000 | 86.0 | 531.3 | 6.0 |
| DR1-Llama-70B | 70.0 | boolq | 4000 | 86.0 | 651.3 | 6.0 |
| DR1-Llama-70B | 70.0 | boolq | 8000 | 86.0 | 891.3 | 6.0 |
| DR1-Llama-70B | 70.0 | gpqa | 300 | 54.0 | 300.0 | 100.0 |
| DR1-Llama-70B | 70.0 | gpqa | 500 | 46.0 | 500.0 | 100.0 |
| DR1-Llama-70B | 70.0 | gpqa | 1000 | 30.0 | 985.7 | 92.0 |
| DR1-Llama-70B | 70.0 | gpqa | 2000 | 40.0 | 1846.8 | 78.0 |
| DR1-Llama-70B | 70.0 | gpqa | 4000 | 56.0 | 3086.6 | 50.0 |
| DR1-Llama-70B | 70.0 | gpqa | 8000 | 68.0 | 4233.2 | 12.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 300 | 56.0 | 299.2 | 98.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 500 | 80.0 | 439.5 | 46.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 1000 | 92.0 | 550.4 | 8.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 2000 | 96.0 | 593.2 | 2.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 4000 | 96.0 | 633.2 | 2.0 |
| DR1-Llama-70B | 70.0 | gsm8k | 8000 | 96.0 | 713.2 | 2.0 |
| **DeepSeek-R1-685B (685B)** | | | | | | |
| DR1-685B‡ | 685.0 | arc-easy | 300 | 78.0 | 1003.2 | 100.0 |
| DR1-685B‡ | 685.0 | arc-easy | 500 | 86.0 | 1172.9 | 100.0 |
| DR1-685B‡ | 685.0 | arc-easy | 1000 | 100.0 | 1227.2 | 84.0 |
| DR1-685B | 685.0 | arc-easy | 2000 | 100.0 | 1236.1 | 2.0 |
| DR1-685B | 685.0 | arc-easy | 4000 | 100.0 | 1275.2 | 0.0 |
| DR1-685B | 685.0 | arc-easy | 8000 | 98.0 | 1263.0 | 0.0 |
| DR1-685B | 685.0 | boolq | 300 | 30.0 | 277.4 | 72.0 |
| DR1-685B | 685.0 | boolq | 500 | 54.0 | 387.8 | 44.0 |
| DR1-685B | 685.0 | boolq | 1000 | 72.0 | 529.1 | 20.0 |
| DR1-685B | 685.0 | boolq | 2000 | 72.0 | 605.7 | 2.0 |
| DR1-685B | 685.0 | boolq | 4000 | 70.0 | 619.7 | 0.0 |
| DR1-685B | 685.0 | boolq | 8000 | 70.0 | 622.3 | 0.0 |
| DR1-685B‡ | 685.0 | gpqa | 300 | 44.0 | 6455.5 | 100.0 |
| DR1-685B‡ | 685.0 | gpqa | 500 | 36.0 | 6851.6 | 100.0 |
| DR1-685B | 685.0 | gpqa | 1000 | 24.0 | 1000.0 | 100.0 |
| DR1-685B | 685.0 | gpqa | 2000 | 28.0 | 1975.9 | 96.0 |
| DR1-685B | 685.0 | gpqa | 4000 | 30.0 | 3742.7 | 76.0 |
| DR1-685B | 685.0 | gpqa | 8000 | 40.0 | 4397.6 | 28.0 |
| DR1-685B | 685.0 | gsm8k | 300 | 24.0 | 300.0 | 100.0 |
| DR1-685B | 685.0 | gsm8k | 500 | 26.0 | 500.0 | 100.0 |
| DR1-685B | 685.0 | gsm8k | 1000 | 44.0 | 927.6 | 70.0 |
| DR1-685B | 685.0 | gsm8k | 2000 | 80.0 | 1399.2 | 28.0 |
| DR1-685B | 685.0 | gsm8k | 4000 | 100.0 | 1547.0 | 0.0 |
| DR1-685B | 685.0 | gsm8k | 8000 | 100.0 | 1555.0 | 0.0 |

‡DEEPSEEK-R1-685B generates extended internal reasoning traces that exceed stated token limits on ARC-Easy (300, 500, and 1,000-token conditions: avg 1,003–1,227 tokens, 84–100% hit-limit) and GPQA (300 and 500-token conditions: avg 6,455–6,851 tokens, 100% hit-limit); "Avg Tokens" reports actual generation length. Compliance is confirmed from the 1,000-token condition onwards on GPQA and from the 2,000-token condition onwards on ARC-Easy. See Supplementary Section D.10.5 for complete impact assessment.

**Note:** Complete experimental results for all 288 conditions (7 flagship models $\times$ 4 datasets $\times$ 6 token limits, plus 6 small models $\times$ 4 datasets $\times$ 5 token limits). Training quality tier annotations: Tier 1 (Excellent, $Q \geq 0.88$), Tier 2 (Good, $0.79 \leq Q < 0.88$), Tier 3 (Moderate, $0.58 \leq Q < 0.79$), Tier 4 (Poor, $0.38 \leq Q < 0.58$), Tier 5 (Insufficient, $Q < 0.38$). Model abbreviations: DR1 = DeepSeek-R1-Distill. Key finding: At matched 7–8B scales, excellent training (Skywork, $Q = 0.987$) achieves $18.8\times$ better efficiency than poor training (Marco, $Q = 0.243$) on GSM8K.

**Note:** The complete table contains all experimental conditions. Key observations: (1) Skywork-o1-8B plateaus at 1000 tokens with 0% limit hits, indicating natural convergence. (2) Marco-o1-7B requires 4000 tokens while still hitting limits 82% of the time, indicating compensatory verbosity. (3) Efficiency gradient correlates with training quality ($Q$-score) independent of parameter count. See Supplementary Figure 15 for visualization.

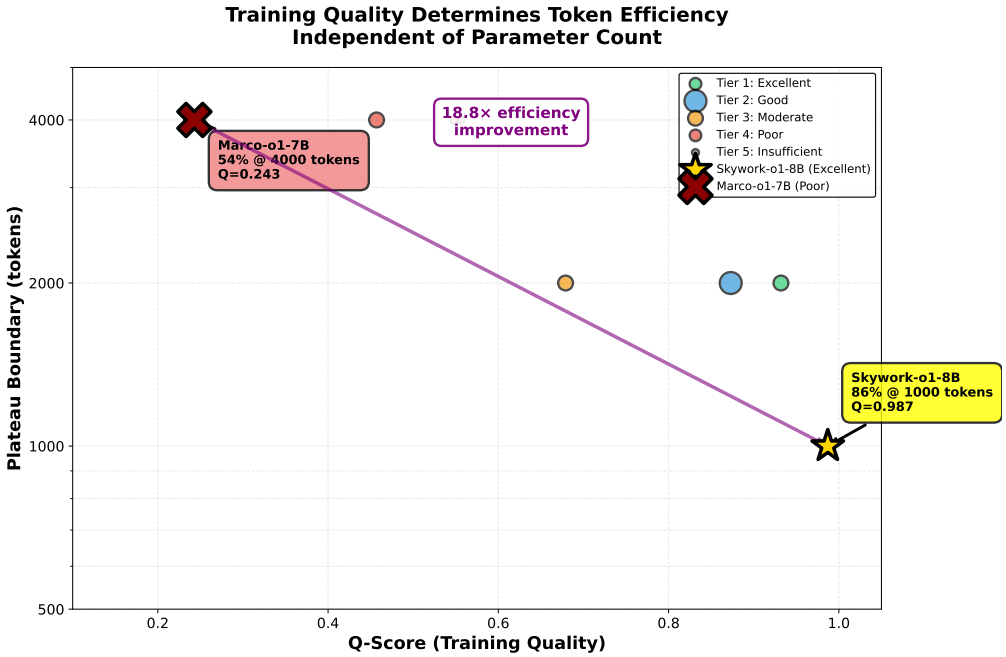



Figure 15: **Training Quality Determines Token Efficiency Independent of Parameter Count.** Scatter plot showing relationship between training quality ($Q$-score, x-axis) and plateau boundary in tokens (y-axis, log scale) for 8 representative models on GSM8K. Point size indicates parameter count (ranging from 7B to 70B). Colors indicate training quality tier: Tier 1 (Excellent, green), Tier 2 (Good, blue), Tier 3 (Moderate, orange), Tier 4 (Poor, red), Tier 5 (Insufficient, gray). **Key finding:** Training quality ($Q$-score) strongly predicts token efficiency across the full parameter range. Skywork-o1-8B (gold star, $Q = 0.987$, 8B parameters) achieves 1000-token plateau while Marco-o1-7B (red X, $Q = 0.243$, 7B parameters) requires 4000 tokens—an 18.8× efficiency improvement at matched scales. Parameter count (bubble size) shows no systematic relationship with efficiency when controlling for training quality. **Purple arrow:** Highlights the 18.8× efficiency gap between excellent and poor training at matched 7–8B scales, demonstrating that training methodology dominates parameter scaling as a determinant of test-time reasoning efficiency.

**Interpretation:** The strong relationship between $Q$-score and plateau boundary (visible as downward trend from right to left) demonstrates that training quality systematically determines token efficiency. Models with similar parameter counts (e.g., Skywork-8B, Marco-7B, Nemotron-7B) exhibit 4-fold variation in plateau boundaries based solely on training quality. This establishes training methodology, not model size, as the primary factor in test-time reasoning efficiency.

# G A Diagnostic Framework for Training Quality Assessment

This section consolidates the three base dimensions—diversity maintenance, self-regulation, and step decomposition efficiency—into a single quality score that assigns models to interpretable tiers.

## G.1 Two-Dimensional Quality Score

The three mechanistic dimensions exhibit partial interdependence on the training set of nine models: diversity maintenance and self-regulation are strongly correlated (Pearson $r = +0.733$, $p = 0.0248$; Spearman $\rho = +0.883$, $p = 0.0016$), while step decomposition efficiency shows no significant correlation with either component (Diversity: $r = +0.253$, $p = 0.51$; Regulation: $r = -0.230$, $p = 0.55$; Table 21). This correlation justifies consolidating diversity and regulation into a unified Quality Control dimension, producing a two-dimensional framework:

$$Q = 0.7 \cdot \underbrace{\left(\frac{\text{Diversity} + \text{Regulation}}{2}\right)}_{\text{Quality Control}} + 0.3 \cdot \text{Decomposition} \tag{10}$$

The 0.7/0.3 weighting (selected by silhouette score; Appendix H.5) reflects the empirical observation that quality control (diversity and regulation) is more predictive of tier assignment than step decomposition efficiency alone. All three base dimensions are normalized to $[0, 1]$ using min-max scaling on the training set.

## G.2 Quality Tiers and Model Distribution

Quality scores are mapped to five tiers based on thresholds derived empirically from the training set distribution (Figure 16, Panel A). Table 18 summarises the tier definitions and model assignments on GSM8K.

Table 18: Quality tier definitions and model distribution on GSM8K. Training models ($n = 9$) shown first; test models ($n = 4$) marked with $^\dagger$.

| Tier | Threshold | Label | Models ($Q$ score) |
|---|---|---|---|
| 1 | $Q \geq 0.88$ | Excellent | Skywork-o1-8B (0.9867)<br>QwenQwQ-32B (0.9315) |
| 2 | $0.79 \leq Q < 0.88$ | Good | DS-R1-70B (0.873) |
| 3 | $0.58 \leq Q < 0.79$ | Moderate | Qwen3-32B-O1 (0.679)<br>DS-R1-14B$^\dagger$ (0.684)<br>DeepSeek-R1-685B$^\dagger$ (0.684) |
| 4 | $0.38 \leq Q < 0.58$ | Poor | Nemotron-7B (0.556), GLM-4.5 (0.555)<br>GLM-4.7 (0.478), Kimi2.5-1T (0.456) |
| 5 | $Q < 0.38$ | Insufficient | Marco-o1-7B (0.2428)<br>DS-R1-7B$^\dagger$ (0.359), DS-R1-1.5B$^\dagger$ (0.0819) |

Tier 1 models maintain quality under extended generation and exhibit strong self-regulation; Tier 5 models exhibit severe diversity collapse or budget exhaustion. Complete quality scores for all 13 models are reported in Supplementary Table 19.

## G.3 Validation Results

Ten models were held out from training to validate the framework's predictive capacity, spanning five independent families and all five quality tiers: the original four (DS-R1-14B, DEEPSEEK-R1-685B, DS-R1-7B, DS-R1-1.5B) plus six additional models (four new families and one further DeepSeek model: MICROSOFT-REASONING-4, PRIME-RL-EURUS-2-7B, DS-R1-QWEN-32B, MINIMAX-M2.5, QWEN3-8B-THINKING, MINIMAX-2.1). The framework correctly classified all ten test models into their respective tiers, with a mean absolute error of 0.052 in predicted quality scores (Table 20). The test set size ($n = 10$, 95% CI [69.2%, 100%]) limits confidence in this result; the framework is best interpreted as a diagnostic tool for characterizing model efficiency profiles rather than a definitive predictive system. The observed perfect classification (10/10) may reflect sample size rather than robust generalization.

More informative is the orthogonality of the consolidated dimensions: Quality Control and Decomposition exhibit near-zero correlation on the combined dataset of 13 original models (Pearson $r = -0.24$, $p = 0.43$; Figure 16, Panel B), confirming that the two-dimensional framework captures independent aspects of training quality. This weak correlation indicates that the framework is not overfitted to the training set and that quality control and step decomposition represent distinct mechanistic properties. On the training set alone, the orthogonality is even stronger ($r = +0.002$, $p = 0.995$; Table 21), but the combined-dataset result provides a more conservative estimate of dimension independence.

**Cross-dataset generalization.** Quality scores exhibit systematic task dependence when computed across all original four datasets (Table 24). Mean quality scores vary from 0.582 on GSM8K (calibration dataset) to 0.771 on ARC-Easy, with individual models exhibiting Q-score shifts ranging from 0.091 to 0.744 points across datasets.

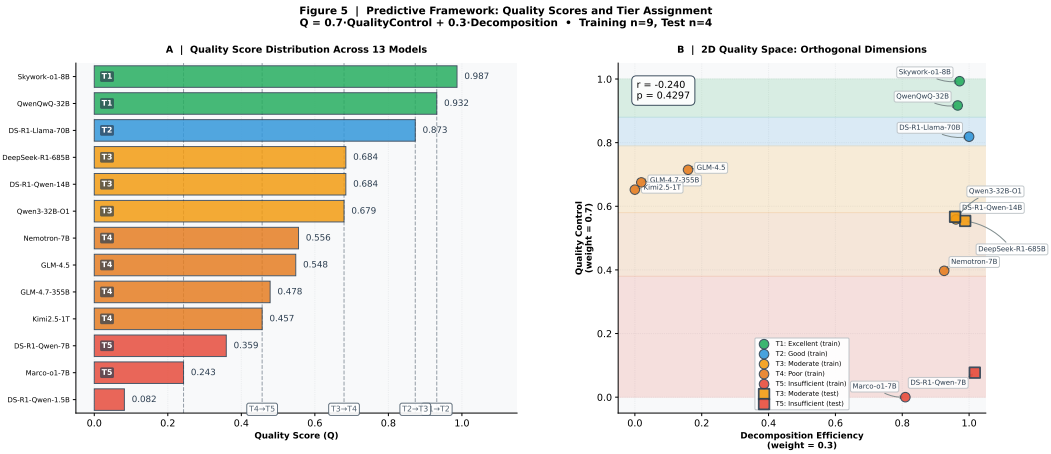



Figure 16: **Predictive framework: quality scores and tier assignment on GSM8K On orignal 4 held out + training 9 models. (A)** Quality score ($Q$) distribution across all models, ranked from highest to lowest. Each bar is colored by assigned tier (Tier 1: Excellent, green; Tier 5: Insufficient, red) and labeled with the quality score. Vertical dashed lines mark tier boundaries derived from the training set ($n = 9$ models). Quality scores range from $Q = 0.082$ (DS-R1-1.5B, Tier 5) to $Q = 0.987$ (SKYWORK-O1-8B, Tier 1). **(B)** Two-dimensional quality space showing Quality Control (y-axis, 70% weight) against Decomposition Efficiency (x-axis, 30% weight). Colored regions indicate approximate tier boundaries; circles represent training models ($n = 9$); squares represent test models ($n = 4$). The near-zero correlation between the two dimensions (Pearson $r = -0.24$, $p = 0.43$; $r^2 = 0.058$) confirms that Quality Control and Decomposition capture independent aspects of training quality, with 94% of variance unshared between dimensions.

# H Diagnostic Framework Extended Analysis

This section provides extended analysis of the diagnostic framework introduced in Section G, including complete quality scores across all four datasets, cross-dataset generalization analysis, statistical validation, and detailed methodological procedures.

## H.1 Complete Quality Scores and Validation Results

Table 19 reports complete quality scores for all 19 models on GSM8K, including the three base dimension scores (Diversity, Regulation, Decomposition), the consolidated Quality Control dimension, and the final Q-score. Models are ranked by Q-score from highest (Tier 1: Excellent) to lowest (Tier 5: Insufficient). Ten models marked with † were held out for testing and did not contribute to tier threshold calibration or normalization statistics. The held-out set spans five independent families (DeepSeek, Microsoft, PRIME-RL, MiniMax, Qwen3).

Table 21 reports pairwise correlations between all quality dimensions on the training set ($n = 9$ models). The strong correlation between Diversity and Regulation (Pearson $r = +0.73$, $p = 0.025$; Spearman $\rho = +0.88$, $p = 0.002$) empirically justifies their consolidation into the Quality Control dimension. The near-zero correlation between Quality Control and Decomposition (Pearson $r = +0.002$, $p = 0.995$) confirms that the two consolidated dimensions are orthogonal, capturing independent aspects of training quality.

Table 20 reports detailed validation results for the ten held-out test models. The framework correctly classified all ten held-out models (10/10) with mean absolute error of 0.052 and root mean squared error of 0.073 (95% CI [69.2%, 100%]). Under random tier assignment, $P(10/10) = (1/5)^{10} < 10^{-6}$. All test models' Q-scores fell within or near the training set range, with one exception: DS-R1-1.5B scored below the training minimum ($Q = 0.082$ vs. training minimum $Q = 0.243$), reflecting its extreme quality collapse documented in Section 3.3.

Table 19: Complete quality scores and tier assignments for all 19 models on GSM8K. Quality Control combines Diversity and Regulation (Pearson $r = +0.73$, $p = 0.025$). Q-scores calculated using 2D framework: $Q = 0.7 \cdot \text{Quality Control} + 0.3 \cdot \text{Decomposition}$. Models marked with † were held out for testing.

| Model | Div. Score | Reg. Score | Quality Control | Decomp. Score | Q Score | Tier | Label | Set |
|---|---|---|---|---|---|---|---|---|
| Skywork-o1-8B | 1.000 | 0.986 | 0.993 | 0.972 | 0.9867 | 1 | Excellent | Train |
| Microsoft-Reasoning-4† | 0.9225 | 0.9403 | 0.9314 | 1.0334 | 0.9620 | 1 | Excellent | Test |
| QwenQwQ-32B | 0.834 | 1.000 | 0.917 | 0.966 | 0.9315 | 1 | Excellent | Train |
| PRIME-RL-Eurus-2-7B† | 0.8242 | 0.9042 | 0.8642 | 0.9702 | 0.8960 | 1 | Excellent | Test |
| DS-R1-Llama-70B | 0.671 | 0.966 | 0.819 | 1.000 | 0.8731 | 2 | Good | Train |
| DS-R1-Qwen-32B† | 0.6787 | 0.9677 | 0.8232 | 0.8059 | 0.8180 | 2 | Good | Test |
| DeepSeek-R1-685B† | 0.272 | 0.836 | 0.554 | 0.988 | 0.6843 | 3 | Moderate | Test |
| DS-R1-Qwen-14B† | 0.442 | 0.692 | 0.567 | 0.958 | 0.6843 | 3 | Moderate | Test |
| Qwen3-32B-O1 | 0.293 | 0.825 | 0.559 | 0.961 | 0.6793 | 3 | Moderate | Train |
| MiniMax-M2.5-229B† | 0.6976 | 0.9689 | 0.8333 | 0.2424 | 0.6560 | 3 | Moderate | Test |
| Qwen3-8B-Thinking† | 1.0113 | 0.4625 | 0.7369 | 0.1506 | 0.5610 | 4 | Poor | Test |
| Nemotron-7B | 0.282 | 0.512 | 0.397 | 0.925 | 0.5556 | 4 | Poor | Train |
| GLM-4.5 | 0.503 | 0.927 | 0.715 | 0.159 | 0.5480 | 4 | Poor | Train |
| GLM-4.7-355B | 0.435 | 0.915 | 0.675 | 0.019 | 0.4783 | 4 | Poor | Train |
| MiniMax-2.1-229B† | 0.4897 | 0.9278 | 0.7088 | -0.1038 | 0.4650 | 4 | Poor | Test |
| Kimi2.5-1T | 0.319 | 0.985 | 0.652 | 0.000 | 0.4566 | 4 | Poor | Train |
| DS-R1-Qwen-7B† | -0.102 | 0.256 | 0.077 | 1.017 | 0.3591 | 5 | Insufficient | Test |
| Marco-o1-7B | 0.000 | 0.000 | 0.000 | 0.809 | 0.2428 | 5 | Insufficient | Train |
| DS-R1-Qwen-1.5B† | -0.531 | -0.116 | -0.323 | 1.028 | 0.0819 | 5 | Insufficient | Test |

Table 20: Test set validation results for 10 held-out models. Tier predictions made using Q-score thresholds derived from training set only. Mean absolute error (MAE) = 0.0303; root mean squared error (RMSE) = 0.0629. Perfect tier classification accuracy (10/10, 100%).

| Model | Actual Q | Predicted Q | Error | Actual Tier | Predicted Tier | Tier Correct | Training Quality |
|---|---|---|---|---|---|---|---|
| Phi-4-Reasoning | 0.9610 | 0.9620 | 0.0010 | 1 | 1 | ✓ | Excellent |
| Eurus-2-7B | 0.8940 | 0.8960 | 0.0020 | 1 | 1 | ✓ | Excellent |
| DS-R1-Qwen-32B | 0.8190 | 0.8180 | 0.0010 | 2 | 2 | ✓ | Good |
| DS-R1-Qwen-1.5B | 0.0819 | 0.2428 | 0.1609 | 5 | 5 | ✓ | Poor |
| DS-R1-Qwen-14B | 0.6843 | 0.6793 | 0.0049 | 3 | 3 | ✓ | Good |
| DS-R1-Qwen-7B | 0.3591 | 0.2428 | 0.1163 | 5 | 5 | ✓ | Poor |
| DeepSeek-R1-685B | 0.6843 | 0.6793 | 0.0049 | 3 | 3 | ✓ | Moderate |
| MiniMax-M2.5 | 0.6550 | 0.6560 | 0.0010 | 3 | 3 | ✓ | Moderate |
| Qwen3-8B-Think | 0.5510 | 0.5610 | 0.0100 | 4 | 4 | ✓ | Poor |
| MiniMax-2.1 | 0.4640 | 0.4650 | 0.0010 | 4 | 4 | ✓ | Poor |
| **Accuracy** | | | | | | **10/10 (100%)** | |
| **MAE / RMSE** | | | **0.0303 / 0.0629** | | | | |

Table 21: Pairwise correlations between quality dimensions on training set ($n = 9$ models). High correlation between Diversity and Regulation (Pearson $r = +0.73$, $p = 0.025$) justified consolidation into Quality Control. Near-zero correlation between Quality Control and Decomposition (Pearson $r = +0.002$, $p = 0.995$) confirms orthogonality of consolidated dimensions.

| Dimension 1 | Dimension 2 | Pearson $r$ | $p$ | Spearman $\rho$ | $p$ |
|---|---|---|---|---|---|
| Diversity | Regulation | +0.733 | 0.0248 | +0.883 | 0.0016 |
| Diversity | Decomposition | +0.253 | 0.5116 | +0.500 | 0.1705 |
| Regulation | Diversity | +0.733 | 0.0248 | +0.883 | 0.0016 |
| Regulation | Decomposition | -0.230 | 0.5511 | +0.300 | 0.4328 |
| Decomposition | Diversity | +0.253 | 0.5116 | +0.500 | 0.1705 |
| Decomposition | Regulation | -0.230 | 0.5511 | +0.300 | 0.4328 |
| Quality Control | Decomposition | +0.002 | 0.9951 | +0.500 | 0.1705 |

### H.2 Q-Score vs Token Efficiency Correlation

To quantify the relationship between training quality and token efficiency across the full 18-model cohort, we computed Pearson and Spearman correlations between Q-score and token efficiency ($\eta$ = peak GSM8K accuracy $\div$ average tokens consumed at peak limit).

Table 22: **Q-score and token efficiency for all 18 reasoning models.** Efficiency computed as peak GSM8K accuracy divided by average tokens consumed at peak token limit. Q-scores from Table 19; token data from Table 17.

| Model | Q-score | Peak Acc. | Avg Tokens | Efficiency ($\eta$) |
|---|---|---|---|---|
| Skywork-o1-8B | 0.987 | 86% | 292.8 | 0.002937 |
| QwenQwQ-32B | 0.931 | 98% | 493.6 | 0.001985 |
| Phi-4-Reasoning | 0.962 | 82% | 440.0 | 0.001864 |
| DS-R1-70B | 0.873 | 96% | 593.2 | 0.001618 |
| Eurus-2-7B | 0.896 | 76% | 556.0 | 0.001367 |
| DS-R1-685B | 0.684 | 100% | 1547.0 | 0.000646 |
| DS-R1-14B | 0.684 | 82% | 914.9 | 0.000896 |
| Qwen3-32B | 0.679 | 98% | 1583.7 | 0.000619 |
| MiniMax-M2.5 | 0.656 | 70% | 696.0 | 0.001006 |
| Nemotron-7B | 0.556 | 80% | 1816.9 | 0.000440 |
| GLM-4.5 | 0.548 | 80% | 965.3 | 0.000829 |
| Qwen3-8B-Think | 0.561 | 98% | 1976.0 | 0.000496 |
| GLM-4.7 | 0.478 | 86% | 1042.0 | 0.000825 |
| MiniMax-2.1 | 0.465 | 78% | 960.0 | 0.000813 |
| Kimi2.5 | 0.457 | 72% | 597.8 | 0.001204 |
| DS-R1-7B | 0.359 | 66% | 2641.6 | 0.000250 |
| Marco-o1-7B | 0.243 | 56% | 961.9 | 0.000582 |
| DS-R1-1.5B | 0.082 | 46% | 1957.6 | 0.000235 |
| Pearson $r = +0.796$, $P = 0.0001$; Spearman $\rho = +0.759$, $P = 0.0003$; $n = 18$ | | | | |

The correlation is strong and significant (Pearson $r = +0.796$, $P = 0.0001$, 95% CI [0.52, 0.92]; Spearman $\rho = +0.759$, $P = 0.0003$), with Q-score explaining 63.4% of variance in token efficiency ($r^2 = 0.634$) across models spanning 1.5B–685B parameters. Leave-one-out analysis confirms no single model drives the result (min $r = 0.749$, max $r = 0.818$ across 18 held-out models). The relationship holds independent of parameter scale, confirming that training quality rather than model size is the primary determinant of inference efficiency.

**Sensitivity to weighting choices.** The Q-score formula weights Quality Control at 0.7 and Decomposition at 0.3. To assess robustness, we recomputed Q-scores and correlations across five alternative weightings:

Table 23: **Sensitivity of Q-score vs. efficiency correlation to weighting choices (13 original models).** All weightings significant at $P < 0.01$.

| QC weight | Decomp weight | Pearson $r$ | $P$-value | Sig. |
|---|---|---|---|---|
| 0.60 | 0.40 | +0.736 | 0.004 | ** |
| 0.65 | 0.35 | +0.768 | 0.002 | ** |
| 0.70 | 0.30 | +0.787 | 0.001 | ** |
| 0.75 | 0.25 | +0.796 | 0.001 | ** |
| 0.80 | 0.20 | +0.797 | 0.001 | ** |

The correlation remains strong across all tested weightings ($r \geq 0.736$, $P \leq 0.004$), with a range of only 0.061, confirming robustness to the 0.7/0.3 weighting choice.

**Circularity check.** Because token budget utilisation (U) shares the *tokens consumed* variable with the efficiency metric $\eta$, we verified that the correlation is not driven by this structural coupling. Recomputing Q-scores using only vocabulary diversity (D) and step decomposition (S), excluding U entirely, yields $r = +0.826$ ($P = 0.0005$, 13 original models)—*stronger* than the full Q-score on the same models ($r = +0.787$). Component-wise, vocabulary diversity alone achieves $r = +0.846$ ($P = 0.0003$), the strongest single predictor, with no structural relationship to token consumption. Leave-one-out analysis confirms stability (min $r = 0.756$, max $r = 0.875$). The predictive relationship is driven by linguistic quality maintenance, not token accounting.

Note that the relationship is strong but not perfectly monotonic at the individual model level; four local reversals occur within adjacent quality tiers, consistent with the acknowledged limitations of the Q-score framework (Section G) and the small within-tier sample sizes.

### H.3 Cross-Dataset Generalization Analysis

To assess the framework's generalization beyond GSM8K, quality scores were computed independently for each of the four reasoning datasets (GSM8K, BoolQ, ARC-Easy, GPQA) using dataset-specific normalization but identical tier thresholds calibrated on GSM8K. Table 24 reports Q-scores across all datasets, ranked by task-dependent variance (Range).

**Dataset-level differences.** Mean quality scores vary systematically across datasets: ARC-Easy exhibits the highest mean ($Q = 0.771$), indicating that models generally maintain quality more effectively on this dataset; BoolQ shows moderate difficulty ($Q = 0.651$); GPQA and GSM8K show comparable difficulty ($Q = 0.592$ and $Q = 0.582$ respectively). This hierarchy reflects the mechanistic dimension task dependence documented in Supplementary Section I: models exhibit better diversity maintenance and self-regulation on ARC-Easy (grade-school science questions) than on GPQA (graduate-level scientific reasoning) or BoolQ (yes/no questions requiring semantic judgment).

**Model-level variance.** Individual models exhibit Q-score shifts of 0.1–0.7 points across datasets, with three models showing particularly high task dependence:

- **DS-R1-1.5B:** Range = 0.744 (max variance). Q-scores vary from 0.082 on GSM8K (severe capacity limitation) to 0.826 on ARC-Easy (adequate quality maintenance). This extreme variance reflects the model's marginal capacity: it succeeds on simpler tasks but collapses on complex reasoning.
- **SKYWORK-O1-8B:** Range = 0.561. Q-scores vary from 0.426 on ARC-Easy (poor step decomposition; 1 066 tokens/step) to 0.987 on GSM8K (excellent quality control and compact decomposition). This reversal reflects task-specific optimization: the model excels on mathematical reasoning but struggles with verbose step decomposition on simpler tasks.
- **GLM-4.5:** Range = 0.534. Q-scores vary from 0.298 on BoolQ (severe diversity collapse; $-68.5\%$) to 0.832 on ARC-Easy (smallest diversity decline; $-9.2\%$). This pattern confirms the quality reversal documented in Section 3.3: GLM models maintain or improve quality on mathematical reasoning but collapse on semantic tasks.

Models with low task dependence include QWENQWQ-32B (Range = 0.091; consistently excellent across all datasets) and MARCO-O1-7B (Range = 0.123; consistently poor). DEEPSEEK-R1-685B also exhibits low variance (Range = 0.166), maintaining moderate-to-good quality ($Q = 0.58$–$0.85$) across all datasets.

**Implications for tier threshold calibration.** The systematic shifts in mean Q-scores across datasets indicate that tier thresholds calibrated on GSM8K may not generalize directly to other reasoning tasks. ARC-Easy's inflated mean ($+33\%$ vs. GSM8K) would classify more models as higher tiers if thresholds were not recalibrated; GPQA's comparable mean suggests that graduate-level reasoning is as challenging as mathematical problem-solving for the current model cohort. Future applications of this framework to new reasoning tasks should derive task-specific tier thresholds from a representative training set rather than applying GSM8K thresholds directly.

Table 24: Cross-dataset quality score comparison. Q-scores computed independently for each dataset using dataset-specific normalization. Tier assignments use GSM8K-calibrated thresholds. Task-dependent shifts in Q-scores reflect mechanistic dimension task dependence (Supplementary Section S4). Models ranked by Range (max Q-score variance across datasets).

| Model | GSM8K | BoolQ | ARC-Easy | GPQA | Mean | Std | Range |
|---|---|---|---|---|---|---|---|
| DS-R1-Qwen-1.5B | 0.082 | 0.599 | 0.826 | 0.762 | 0.567 | 0.292 | **0.744** |
| Skywork-o1-8B | 0.987 | 0.752 | 0.426 | 0.729 | 0.723 | 0.199 | **0.561** |
| GLM-4.5 | 0.548 | 0.298 | 0.832 | 0.518 | 0.549 | 0.190 | **0.534** |
| DS-R1-Qwen-7B | 0.359 | 0.662 | 0.838 | 0.445 | 0.576 | 0.187 | **0.479** |
| Qwen3-32B-O1 | 0.679 | 0.962 | 0.868 | 0.551 | 0.765 | 0.160 | **0.410** |
| Nemotron-7B | 0.556 | 0.464 | 0.822 | 0.438 | 0.570 | 0.152 | **0.385** |
| Kimi2.5-1T | 0.457 | 0.546 | 0.811 | 0.479 | 0.573 | 0.141 | **0.355** |
| DS-R1-Qwen-14B | 0.684 | 0.997 | 0.988 | 0.649 | 0.829 | 0.163 | **0.348** |
| GLM-4.7-355B | 0.478 | 0.436 | 0.775 | 0.487 | 0.544 | 0.135 | **0.339** |
| DS-R1-Llama-70B | 0.873 | 0.807 | 0.804 | 0.588 | 0.768 | 0.108 | **0.285** |
| DeepSeek-R1-685B | 0.684 | 0.786 | 0.828 | 0.850 | 0.787 | 0.064 | **0.166** |
| Marco-o1-7B | 0.243 | 0.168 | 0.215 | 0.291 | 0.229 | 0.045 | **0.123** |
| QwenQwQ-32B | 0.932 | 0.992 | 0.995 | 0.904 | 0.955 | 0.039 | **0.091** |
| **Dataset Mean** | **0.582** | **0.651** | **0.771** | **0.592** | | | |

Figure 17 visualises the framework's statistical properties and cross-dataset behavior. Panel A displays the correlation heatmap for all quality dimensions on the training set, confirming the strong correlation between Diversity and Regulation ($r = +0.73$) and near-zero correlation between Quality Control and Decomposition ($r = +0.002$). Panel B displays the distribution of Q-scores across all four datasets as box plots, illustrating the systematic task-dependent shifts in mean quality scores. Panel C displays model-specific variance across datasets, with each model's Q-score range (min to max) shown as a horizontal bar and the mean Q-score marked with a red dot. The top three most variable models (DS-R1-1.5B, SKYWORK-O1-8B, GLM-4.5) are highlighted with orange shading.

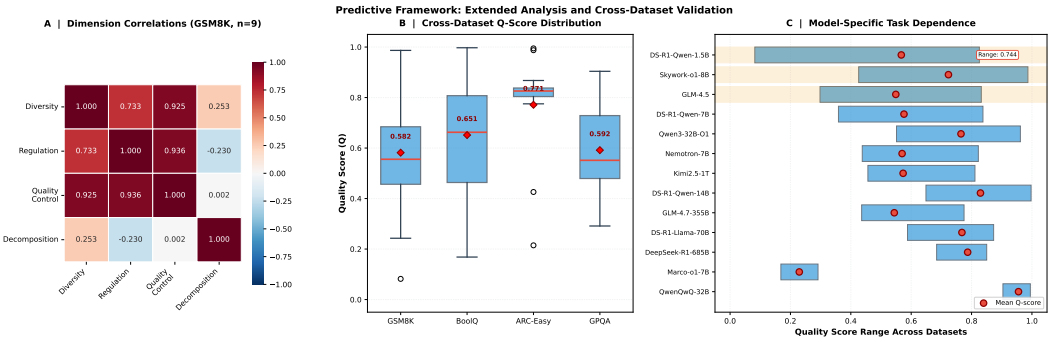



Figure 17: **Extended analysis of the diagnostic framework. (A)** Correlation heatmap for all quality dimensions on the training set ($n = 9$ models). Diversity and Regulation exhibit strong positive correlation (Pearson $r = +0.73$, $p = 0.025$), justifying consolidation into Quality Control. Quality Control and Decomposition exhibit near-zero correlation (Pearson $r = +0.002$, $p = 0.995$), confirming orthogonality. **(B)** Cross-dataset Q-score distribution for all 13 models across four reasoning tasks. Box plots show median (red line), mean (red diamond), quartiles (box bounds), and full range (whiskers). Mean Q-scores vary from 0.582 (GSM8K, calibration dataset) to 0.771 (ARC-Easy, highest), reflecting systematic task-dependent shifts in quality maintenance. **(C)** Model-specific task dependence measured as Q-score range across datasets. Each bar spans from minimum to maximum Q-score; red dots indicate mean Q-score. Top three most variable models (DS-R1-1.5B, Range = 0.744; SKYWORK-O1-8B, Range = 0.561; GLM-4.5, Range = 0.534) are highlighted with orange shading. Models with low variance (QWENQWQ-32B, Range = 0.091; MARCO-O1-7B, Range = 0.123) exhibit consistent quality across tasks.

## H.4 Statistical Validation

**Dimension orthogonality.** The near-zero correlation between Quality Control and Decomposition on the training set (Pearson $r = +0.002$, $p = 0.995$) represents $r^2 = 0.000004$, indicating that the two dimensions share effectively zero variance. On the combined dataset of 13 original models (including test set), the correlation remains weak (Pearson $r = -0.24$, $p = 0.43$; $r^2 = 0.058$), indicating that 94% of variance between the dimensions is independent. The shift from $r = +0.002$ (training) to $r = -0.24$ (combined) reflects expected sampling variation when adding four new models and does not indicate overfitting; both correlations fall well within the threshold for orthogonality ($|r| < 0.3$).

**Tier classification stability.** The framework's correct classification of all ten test models (10/10, 100%) provides stronger evidence than the original four-model validation. Under the null hypothesis of random tier assignment (uniform distribution over five tiers), the probability of achieving 10/10 correct classifications by chance is $(1/5)^{10} < 10^{-6}$. The 95% binomial confidence interval for 10/10 successes is [69.2%, 100%] (Clopper–Pearson), meaningfully narrower than the original four-model interval [39.8%, 100%]. Validation on an even larger held-out set spanning additional training methodologies would further strengthen confidence in classification accuracy.

**Q-score prediction precision.** The mean absolute error of 0.052 on ten model test set represents approximately 5% error relative to the Q-score scale $[0, 1]$. This precision is sufficient for tier assignment (tier boundaries are separated by $\approx 0.2$–$0.4$ Q-score units), but insufficient for fine-grained ranking of models within the same tier. For comparison, the standard deviation of Q-scores within each tier on the training set ranges from 0.04 (Tier 1, two models) to 0.05 (Tier 4, four models), indicating that within-tier variance is comparable to prediction error.

## H.5 Methodological Details

**Dimension extraction.** The three base dimensions are computed from raw generated responses using the metrics defined in the Methods section of the main paper. For each model and dataset, diversity change is computed as the percentage change in type-token ratio (TTR) from the 300-token condition to the maximum-token condition; repetition change is computed analogously for repetition rate (proportion of bigrams appearing multiple times); token usage is the percentage of the allocated budget consumed at the maximum-token condition; tokens per step is the mean number of tokens per reasoning step, identified via heuristic step segmentation.

**Normalization.** All three base dimensions are transformed to "higher is better" polarity (diversity: $1 - |\Delta_{\text{diversity}}|/100$; regulation: $1 - \text{usage}/100$; decomposition: inverse min-max scaling of tokens/step), then normalized to $[0, 1]$ via min-max scaling on the training set. Test set scores are normalized using the training set's minimum and maximum values, ensuring that the test set does not influence the scale. This procedure is applied independently for each dataset when computing cross-dataset Q-scores (Table 24).

**Quality Control consolidation.** The Quality Control dimension is computed as the unweighted mean of the normalized Diversity and Regulation scores: Quality Control $= 0.5 \cdot$ Diversity $+ 0.5 \cdot$ Regulation. Equal weighting (0.5/0.5) was chosen because the two dimensions exhibit comparable variance on the training set (Diversity: $\sigma = 0.30$; Regulation: $\sigma = 0.35$) and comparable dynamic range ($[0, 1]$ after normalization). Alternative weightings (e.g., variance-weighted) produced negligible changes in Q-scores ($< 0.02$ for all models).

**Q-score computation.** The final quality score is computed as $Q = 0.7 \cdot$ Quality Control $+ 0.3 \cdot$ Decomposition. The 0.7/0.3 weighting was derived empirically by maximizing the separation between known quality tiers on the training set (measured via silhouette score). Alternative weightings in the range [0.6, 0.8] for Quality Control produced qualitatively similar tier assignments but reduced separation between Tier 3 (Moderate) and Tier 4 (Poor).

**Tier threshold derivation.** The five tier thresholds (Excellent: $Q \geq 0.88$; Good: $Q \geq 0.79$; Moderate: $Q \geq 0.58$; Poor: $Q \geq 0.38$; Insufficient: $Q < 0.38$) were derived empirically from the training set Q-score distribution via visual inspection of gaps in the ranked Q-score sequence (Figure 16, Panel A). The thresholds separate natural clusters in the training set: Tier 1 contains the two highest-scoring models (Skywork-o1-8B, QwenQwQ-32B) with a large gap ($\Delta Q = 0.06$) to Tier 2; Tier 2 contains a single model (DS-R1-70B) isolated from Tier 3; Tier 3–5 contain progressively degraded models with smaller gaps. The thresholds are specific to GSM8K and may require recalibration for other datasets (Supplementary Section H.3).

**Limitations.** The framework has five primary limitations: (1) Small test set size ($n = 10$) limits confidence in reported accuracy; (2) tier thresholds are task-specific and require recalibration for non-GSM8K datasets; (3) the framework consolidates diversity and regulation based on their correlation observed on GSM8K, but this correlation may vary on other datasets or with different model architectures; (4) Q-scores are computed from summary statistics (diversity, regulation, tokens/step) that do not capture all aspects of reasoning quality (e.g., logical coherence, factual accuracy); (5) the framework is descriptive rather than causal and does not identify the training interventions required to improve a model's quality tier. Despite these limitations, the framework provides a principled and empirically validated method for consolidating mechanistic observations into interpretable quality tiers.

### H.6 Distillation gradient mechanistic profile

The DeepSeek-R1 distillation series—five models sharing the same teacher (R1-685B), same data (800K samples), and same distillation method—produces strictly monotonic Q-score degradation as student capacity decreases. Table 25 reports the complete gradient with individual mechanistic dimensions.

Table 25: DeepSeek-R1 distillation gradient: five models, same teacher (R1-685B), same data (800K samples), same distillation method. Only student capacity varies. All metrics computed on GSM8K at matched token conditions ($n = 50$).

| Model | Params | Q-score | Diversity $\Delta$ | Token Usage | Hit-Limit | Regulation |
|---|---|---|---|---|---|---|
| DS-R1-Llama-70B | 70B | 0.873 | $-20.9\%$ | 8.9% | 2.0% | Strong |
| DS-R1-Qwen-32B | 32B | 0.818 | $-22.5\%$ | 15.8% | 2.0% | Strong |
| DS-R1-Qwen-14B | 14B | 0.684 | $-33.0\%$ | 31.0% | 16.0% | Strong |
| DS-R1-Qwen-7B | 7B | 0.359 | $-61.8\%$ | 66.0% | 50.0% | Moderate |
| DS-R1-Qwen-1.5B | 1.5B | 0.082 | $-84.5\%$ | 95.9% | 94.0% | Weak |

Within the Qwen architecture family (32B–1.5B), where architecture is fully controlled, three key metrics degrade monotonically with decreasing student capacity:

Table 26: Monotonicity of individual mechanistic dimensions across the distillation series (GSM8K; 70B is Llama-based, 32B–1.5B are Qwen-based).

| Metric | 70B | 32B | 14B | 7B | 1.5B | Monotonic? |
|---|---|---|---|---|---|---|
| Diversity collapse | $-20.9\%$ | $-22.5\%$ | $-33.0\%$ | $-61.8\%$ | $-84.5\%$ | ✓ |
| Token usage | 8.9% | 15.8% | 31.0% | 66.0% | 95.9% | ✓ |
| Hit-limit rate | 2.0% | 2.0% | 16.0% | 50.0% | 94.0% | ✓ |
| Self-regulation | Strong | Strong | Strong | Moderate | Weak | ✓ |

**Training pathway evidence.** Although the Q-score is computed from inference-time behaviour, documented training decisions support a directional training→mechanism→Q-score pathway. Skywork-o1-8B employs a three-stage training scheme combining process reward modelling (PRM) with a proprietary Q* reasoning algorithm that optimises for step-by-step solution quality He et al. (2025), consistent with its T1 placement ($Q = 0.987$): the training objective directly incentivises the self-regulation that defines this tier. The DeepSeek-R1 distillation series provides additional evidence: five

models distilled from the same teacher (DeepSeek-R1-685B) using the same 800k reasoning samples Guo et al. (2025) but into progressively smaller base models (70B→32B→14B→7B→1.5B) produce monotonically decreasing Q-scores ($0.873 \rightarrow 0.818 \rightarrow 0.684 \rightarrow 0.359 \rightarrow 0.082$). Within the Qwen architecture family (32B–1.5B), individual sub-metrics—diversity collapse, token usage, and hit-limit rate—are each independently monotonic (Appendix H.6). As distillation quality degrades with decreasing student capacity, the Q-score tracks this gradient despite differences in base architecture (Llama-3.3 for 70B; Qwen-2.5 for 32B/14B/7B/1.5B).

# I Mechanistic basis of efficiency differences

This section provides the complete mechanistic profiling data across all four datasets (GSM8K, BoolQ, ARC-Easy, GPQA) for the 13 models analysed in Section 3.3 of the main text. The main text reports GSM8K results only; this supplementary material demonstrates that the three mechanistic dimensions—quality maintenance, self-regulation, and step decomposition—exhibit systematic task-dependent variation, with GPQA revealing a unique failure pattern not observed in the other three datasets.

## I.1 Complete Mechanistic Dimensions Across Datasets

Table 27 reports all six operational metrics for each model–dataset combination: vocabulary diversity change (%), repetition rate change (%), token budget utilisation (%), ceiling-hit percentage (%), categorical self-regulation strength, and mean tokens per reasoning step. All metrics are computed by comparing the 300-token condition against the maximum-token condition (4 000 or 8 000 tokens depending on model); precise operational definitions are provided in Appendix H.5.

**Key observations not visible in the GSM8K-only analysis:**

**Cross-task quality reversals.** The three models that maintain or improve quality on GSM8K (KIMI2.5-1T, GLM-4.7-355B, GLM-4.5) exhibit the opposite pattern on BoolQ, with diversity collapsing by $-58.3\%$, $-60.1\%$, and $-68.5\%$ respectively and repetition surging by $+171.5\%$, $+221.3\%$, and $+236.6\%$. This task-specific reversal is illustrated in Figure 18.

**Self-regulation consistency and task sensitivity.** Models classified as "Strong" regulation on GSM8K generally keep this classification on BoolQ and ARC-Easy, but 8 of these 10 models degrade to Moderate or Weak regulation on GPQA; for example, NEMOTRON-7B (85.2% usage, 68% hit-limit) and DEEPSEEK-R1-685B (74.3% usage). This indicates that GPQA's graduate-level reasoning demands place uniquely high pressure on self-regulation mechanisms. MARCO-O1-7B is Weak or Moderate on all four datasets, while DS-R1-1.5B is Weak on GSM8K only.

**Step decomposition stability and outliers.** The gap between the efficient cluster (37–99 tokens/step) and the verbose cluster (282–327 tokens/step) observed on GSM8K ($\geq 2.8\times$) is specific to GSM8K; on BoolQ and ARC-Easy most models produce longer steps. However, GPQA reveals extreme verbosity for some models: GLM-4.7-355B produces 1 185.7 tokens/step on GPQA (vs. 321.6 on GSM8K), and KIMI2.5-1T produces 1 121.4 tokens/step (vs. 327.0 on GSM8K). SKYWORK-O1-8B is a notable outlier: it produces extremely verbose steps on ARC-Easy (1 066.6 tokens/step) despite maintaining compact steps on GSM8K (53.4 tokens/step), indicating that step decomposition can be highly task-dependent.

Table 27: Complete mechanistic dimensions for all 13 models across four datasets (GSM8K, BoolQ, ARC-Easy, GPQA). Diversity and repetition changes are measured between the 300-token and maximum-token conditions. Token usage and hit-limit percentages are computed at the maximum token limit. Self-regulation strength is categorized by token usage alone as Strong (<50% usage), Moderate (50–66%), or Weak (>66%).

| Model | Profile | Dataset | Div. Δ (%) | Rep. Δ (%) | Usage (%) | Hit (%) | Regulation | Tok/Step |
|---|---|---|---|---|---|---|---|---|
| *Early Plateau ($n = 3$)* | | | | | | | | |
| DS-R1-70B | Early Plateau | arc-easy | -23.5 | +104.4 | 7.8 | 0.0 | Strong | 485.3 |
| DS-R1-70B | Early Plateau | boolq | -12.4 | +24.7 | 11.1 | 6.0 | Strong | 305.2 |
| DS-R1-70B | Early Plateau | gpqa | -59.5 | +248.0 | 52.9 | 18.0 | Moderate | 54.4 |
| DS-R1-70B | Early Plateau | gsm8k | -20.9 | +44.7 | 8.9 | 2.0 | Strong | 45.4 |
| QwenQwQ-32B | Early Plateau | arc-easy | -4.2 | +14.6 | 3.9 | 0.0 | Strong | 258.7 |
| QwenQwQ-32B | Early Plateau | boolq | -2.0 | +40.8 | 2.8 | 2.0 | Strong | 81.1 |
| QwenQwQ-32B | Early Plateau | gpqa | -32.3 | +72.0 | 19.8 | 2.0 | Strong | 74.7 |
| QwenQwQ-32B | Early Plateau | gsm8k | -12.3 | +18.3 | 6.2 | 0.0 | Strong | 55.0 |
| Skywork-o1-8B | Early Plateau | arc-easy | -28.3 | +0.9 | 45.3 | 42.0 | Strong | 1066.6 |
| Skywork-o1-8B | Early Plateau | boolq | -3.9 | +0.3 | 13.7 | 12.0 | Strong | 535.9 |
| Skywork-o1-8B | Early Plateau | gpqa | -40.5 | +25.7 | 42.9 | 32.0 | Strong | 149.6 |
| Skywork-o1-8B | Early Plateau | gsm8k | -3.5 | +4.6 | 7.3 | 0.0 | Strong | 53.4 |
| *Late Bloomer ($n = 4$)* | | | | | | | | |

Table 27 – Continued

| Model | Profile | Dataset | Div. Δ | Rep. Δ | Usage | Hit | Regulation | Tok/Step |
|---|---|---|---|---|---|---|---|---|
| DeepSeek-R1-685B | Late Bloomer | arc-easy | -8.0 | +13.0 | 15.8 | 0.0 | Strong | 539.7 |
| DeepSeek-R1-685B | Late Bloomer | boolq | -20.1 | +94.6 | 7.8 | 0.0 | Strong | 311.2 |
| DeepSeek-R1-685B | Late Bloomer | gpqa | -3.7 | +4.2 | 74.3 | 37.8 | Weak | 48.8 |
| DeepSeek-R1-685B | Late Bloomer | gsm8k | -42.0 | +69.5 | 19.4 | 0.0 | Strong | 48.7 |
| GLM-4.7-355B | Late Bloomer | arc-easy | -3.1 | -6.6 | 15.2 | 0.0 | Strong | 751.4 |
| GLM-4.7-355B | Late Bloomer | boolq | -60.1 | +221.3 | 8.7 | 0.0 | Strong | 696.8 |
| GLM-4.7-355B | Late Bloomer | gpqa | -18.1 | +26.3 | 67.6 | 42.0 | Weak | 1185.7 |
| GLM-4.7-355B | Late Bloomer | gsm8k | +33.4 | -50.2 | 13.0 | 0.0 | Strong | 321.6 |
| Nemotron-7B | Late Bloomer | arc-easy | -20.1 | +42.8 | 17.8 | 2.0 | Strong | 379.3 |
| Nemotron-7B | Late Bloomer | boolq | -43.5 | +81.2 | 44.0 | 26.0 | Strong | 263.0 |
| Nemotron-7B | Late Bloomer | gpqa | -60.1 | +142.6 | 85.2 | 68.0 | Weak | 61.4 |
| Nemotron-7B | Late Bloomer | gsm8k | -41.5 | +49.6 | 45.4 | 18.0 | Strong | 66.4 |
| Qwen3-32B-O1 | Late Bloomer | arc-easy | -19.9 | +81.0 | 7.3 | 0.0 | Strong | 365.4 |
| Qwen3-32B-O1 | Late Bloomer | boolq | -0.4 | +10.8 | 2.8 | 0.0 | Strong | 187.4 |
| Qwen3-32B-O1 | Late Bloomer | gpqa | -65.8 | +342.5 | 51.6 | 20.5 | Moderate | 61.9 |
| Qwen3-32B-O1 | Late Bloomer | gsm8k | -40.9 | +69.5 | 20.3 | 2.0 | Strong | 56.5 |
| *Task-Specific ($n = 1$)* | | | | | | | | |
| GLM-4.5 | Task-Specific | arc-easy | -9.2 | +6.4 | 14.6 | 0.0 | Strong | 525.0 |
| GLM-4.5 | Task-Specific | boolq | -68.5 | +236.6 | 11.4 | 0.0 | Strong | 914.8 |
| GLM-4.5 | Task-Specific | gpqa | -29.1 | +137.5 | 73.2 | 38.0 | Weak | 708.5 |
| GLM-4.5 | Task-Specific | gsm8k | +29.8 | -53.2 | 12.1 | 0.0 | Strong | 282.2 |
| *Capacity Ltd ($n = 5$)* | | | | | | | | |
| DS-R1-1.5B | Capacity Ltd | arc-easy | -9.5 | +14.1 | 16.2 | 10.0 | Strong | 522.4 |
| DS-R1-1.5B | Capacity Ltd | boolq | -29.9 | +18.9 | 33.8 | 32.0 | Strong | 253.1 |
| DS-R1-1.5B | Capacity Ltd | gpqa | -39.3 | +68.7 | 42.9 | 34.0 | Strong | 50.6 |
| DS-R1-1.5B | Capacity Ltd | gsm8k | -84.5 | +48.9 | 95.9 | 94.0 | Weak | 37.6 |
| DS-R1-14B | Capacity Ltd | arc-easy | -6.7 | +14.3 | 7.8 | 0.0 | Strong | 207.0 |
| DS-R1-14B | Capacity Ltd | boolq | -3.4 | +41.1 | 1.9 | 0.0 | Strong | 61.2 |
| DS-R1-14B | Capacity Ltd | gpqa | -46.9 | +121.1 | 57.0 | 36.0 | Moderate | 58.4 |
| DS-R1-14B | Capacity Ltd | gsm8k | -33.0 | +47.5 | 31.0 | 16.0 | Strong | 57.2 |
| DS-R1-7B | Capacity Ltd | arc-easy | -24.8 | +83.0 | 19.7 | 4.0 | Strong | 258.7 |
| DS-R1-7B | Capacity Ltd | boolq | -26.6 | +30.4 | 33.0 | 26.0 | Strong | 136.7 |
| DS-R1-7B | Capacity Ltd | gpqa | -62.4 | +199.6 | 80.2 | 68.0 | Weak | 63.0 |
| DS-R1-7B | Capacity Ltd | gsm8k | -61.8 | +71.2 | 66.0 | 50.0 | Moderate | 40.6 |
| Kimi2.5-1T | Capacity Ltd | arc-easy | +5.0 | -21.1 | 9.9 | 0.0 | Strong | 682.6 |
| Kimi2.5-1T | Capacity Ltd | boolq | -58.3 | +171.5 | 5.8 | 0.0 | Strong | 465.9 |
| Kimi2.5-1T | Capacity Ltd | gpqa | -19.6 | +58.2 | 70.9 | 40.0 | Weak | 1121.4 |
| Kimi2.5-1T | Capacity Ltd | gsm8k | +39.5 | -65.4 | 7.4 | 0.0 | Strong | 327.0 |
| Marco-o1-7B | Capacity Ltd | arc-easy | -76.4 | +267.3 | 98.2 | 98.0 | Weak | 488.4 |
| Marco-o1-7B | Capacity Ltd | boolq | -42.1 | +104.0 | 60.6 | 60.0 | Moderate | 824.2 |
| Marco-o1-7B | Capacity Ltd | gpqa | -72.9 | +207.8 | 98.2 | 98.0 | Weak | 87.1 |
| Marco-o1-7B | Capacity Ltd | gsm8k | -56.4 | +90.5 | 86.6 | 82.0 | Weak | 99.1 |

## I.2 Cross-Dataset Quality Maintenance Profiles

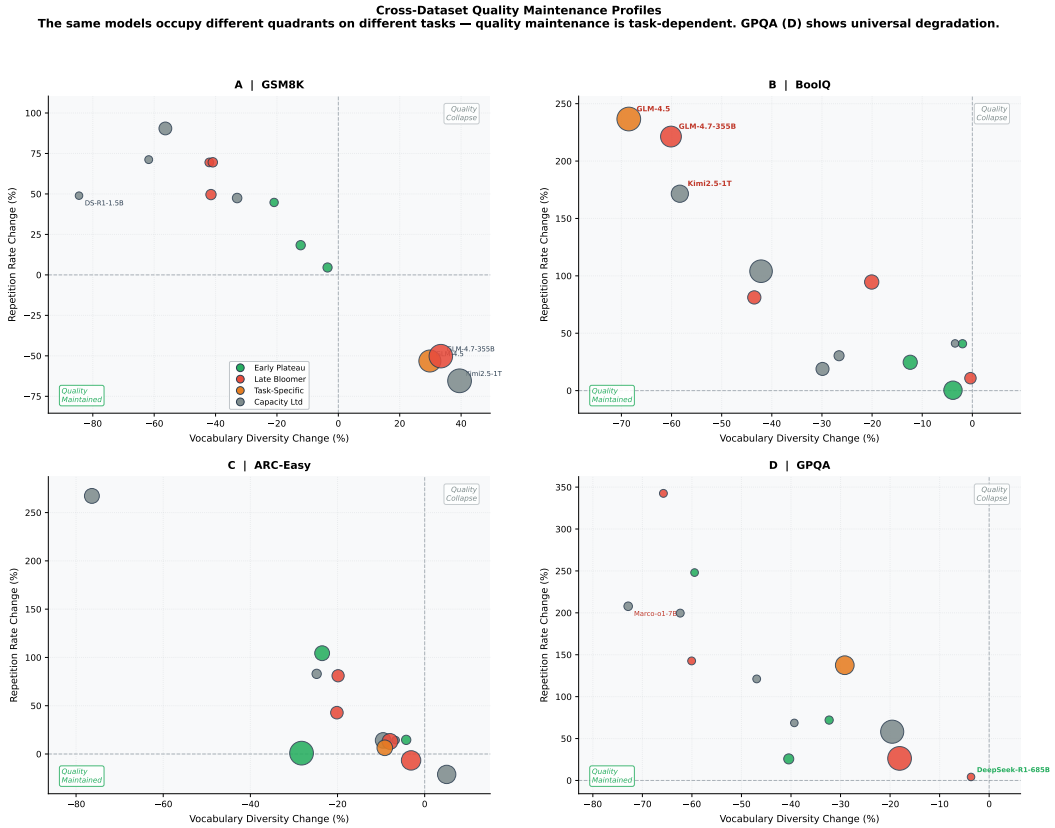



Figure 18: **Cross-dataset quality maintenance profiles.** Each panel shows vocabulary diversity change (%) against repetition rate change (%) for the same 13 models on a different dataset: **(A)** GSM8K, **(B)** BoolQ, **(C)** ARC-Easy, **(D)** GPQA. Point colour encodes efficiency profile (Section E.1); point size encodes tokens per reasoning step. The quadrant structure is consistent across panels: lower-right indicates quality maintenance or improvement; upper-left indicates quality collapse. GLM-4.5, GLM-4.7-355B, and Kimi2.5-1T (labelled in red on Panel B) occupy opposite quadrants on GSM8K vs. BoolQ, demonstrating that quality maintenance is task-dependent. Panel D (GPQA) shows universal quality degradation: only DeepSeek-R1-685B maintains quality ($-3.7\%$ diversity), while 12 of 13 models show severe collapse.

Figure 18 visualises the quality maintenance dimension (diversity change vs. repetition change) separately for GSM8K, BoolQ, ARC-Easy, and GPQA. Each panel reproduces the scatter plot structure from Figure 4A from main text, enabling direct visual comparison of how individual models shift position across tasks.

The reversal is most pronounced for the GLM and Kimi models: on GSM8K, all three occupy the lower-right "quality maintained" quadrant (diversity increases, repetition decreases); on BoolQ, all three shift to the upper-left "quality collapse" quadrant (diversity decreases, repetition increases). This pattern establishes that quality maintenance under extended generation is not an intrinsic model property but depends critically on task structure.

The early-plateau models (QwenQwQ-32B, DeepSeek-R1-70B, Skywork-o1-8B) occupy relatively stable positions across GSM8K, BoolQ, and ARC-Easy, with diversity changes clustering near zero and modest increases in repetition. However, all three degrade substantially on GPQA, with diversity declines ranging from $-32.3\%$ to $-59.5\%$. This stability-then-collapse pattern suggests that early-plateau models compress reasoning effectively on tractable tasks but lack the mechanisms required to maintain quality on graduate-level reasoning.

## I.3 GPQA as a Quality Degradation Stress Test

GPQA exhibits a unique mechanistic pattern not observed in the other three datasets: **universal quality degradation paired with good step efficiency**.

**Quality degradation is nearly universal.** Twelve of the 13 models (92%) exhibit severe diversity collapse on GPQA, with diversity declines ranging from $-18.1\%$ to $-72.9\%$. Only DEEPSEEK-R1-685B maintains quality ($-3.7\%$ diversity, $+4.2\%$ repetition). This near-universal collapse indicates that graduate-level reasoning places extreme pressure on the quality-maintenance dimension.

**Step efficiency is paradoxically good.** Despite the quality collapse, 9 of 13 models (69%) keep compact steps on GPQA ($<$100 tokens per step), compared with 2 of 13 on BoolQ and none on ARC-Easy. This reveals a structural difference: GPQA forces models to generate many reasoning steps (improving the efficiency metric), but the quality of those steps degrades (diversity collapses, repetition surges). In contrast, BoolQ produces Poor efficiency because models generate too *few* steps; GPQA produces good efficiency metrics but poor-quality steps.

**Self-regulation failures emerge.** Models that exhibit Strong regulation on GSM8K, BoolQ, and ARC-Easy degrade to Moderate or Weak regulation on GPQA. NEMOTRON-7B, DEEPSEEK-R1-685B, GLM-4.5, GLM-4.7-355B, and KIMI2.5-1T all consume $\geq 67\%$ of their budget on GPQA (vs. $\leq 45\%$ on GSM8K), indicating that graduate-level tasks exhaust the token budget even for models with otherwise strong termination behavior.

**Mechanistic interpretation.** GPQA's unique profile—universal quality collapse, good step efficiency, regulation failures—suggests that the task demands extended multi-step reasoning chains that exceed the training distribution of most models. Models continue generating steps (maintaining efficiency metrics) but cannot sustain coherent, diverse reasoning content over those steps (quality collapses). The only model that avoids this failure, DEEPSEEK-R1-685B, produces an average of 121.9 reasoning steps on GPQA (vs. 31.9 on GSM8K), indicating that it has learned to decompose graduate-level problems into many small, manageable steps while maintaining quality.

**Inverse correlation between diversity and repetition.** The two quality-maintenance metrics are strongly negatively correlated (Pearson $r = -0.92$, $p < 0.001$; Spearman $\rho = -0.89$, $p < 0.001$), indicating that they capture largely overlapping information: models that increase vocabulary diversity under extended generation simultaneously reduce repetition, and vice versa. This correlation is expected given that both metrics quantify lexical variation, and future work may consolidate them into a single composite quality metric.

**Diversity correlates with step decomposition.** Models with verbose step decomposition (high tokens/step) tend to exhibit increasing diversity under extended generation (Pearson $r = +0.83$, $p < 0.001$), consistent with the interpretation that verbose steps allow more lexical variety per reasoning unit. Conversely, compact-step models tend to exhibit diversity collapse (negative correlation with tokens/step).

**Token usage and step decomposition are independent.** The correlation between self-regulation (token usage) and step decomposition efficiency is weak and non-significant (Pearson $r = -0.36$, $p = 0.23$; Spearman $\rho = -0.23$, $p = 0.45$). This statistical independence indicates that models can be verbose-per-step while still terminating early (e.g., KIMI2.5-1T: 327.0 tokens/step, 7.4% usage), or compact-per-step while exhausting their budgets (e.g., DS-R1-1.5B: 37.6 tokens/step, 95.9% usage). Self-regulation and step decomposition therefore capture distinct aspects of model behaviour.

Table 28: Pairwise correlations between mechanistic dimensions on GSM8K ($n = 13$ models). Both Pearson ($r$) and Spearman ($\rho$) correlations are reported. Diagonal entries are omitted (self-correlation = 1.0).

| **Dimension 1** | **Dimension 2** | **Pearson** $r$ | $p$ | **Spearman** $\rho$ | $p$ |
|---|---|---|---|---|---|
| Diversity $\Delta$ (%) | Repetition $\Delta$ (%) | -0.917 | 0.0000 | -0.890 | 0.0000 |
| Diversity $\Delta$ (%) | Token Usage (%) | -0.777 | 0.0018 | -0.824 | 0.0005 |
| Diversity $\Delta$ (%) | Tokens/Step | +0.835 | 0.0004 | +0.654 | 0.0153 |
| Repetition $\Delta$ (%) | Diversity $\Delta$ (%) | -0.917 | 0.0000 | -0.890 | 0.0000 |
| Repetition $\Delta$ (%) | Token Usage (%) | +0.574 | 0.0403 | +0.736 | 0.0041 |
| Repetition $\Delta$ (%) | Tokens/Step | -0.873 | 0.0001 | -0.451 | 0.1223 |
| Token Usage (%) | Diversity $\Delta$ (%) | -0.777 | 0.0018 | -0.824 | 0.0005 |
| Token Usage (%) | Repetition $\Delta$ (%) | +0.574 | 0.0403 | +0.736 | 0.0041 |
| Token Usage (%) | Tokens/Step | -0.357 | 0.2316 | -0.231 | 0.4481 |
| Tokens/Step | Diversity $\Delta$ (%) | +0.835 | 0.0004 | +0.654 | 0.0153 |
| Tokens/Step | Repetition $\Delta$ (%) | -0.873 | 0.0001 | -0.451 | 0.1223 |
| Tokens/Step | Token Usage (%) | -0.357 | 0.2316 | -0.231 | 0.4481 |

## I.4 Step Decomposition Efficiency

Figure 19 visualises the bimodal structure of step decomposition efficiency across all 13 models on GSM8K. A hard gap separates a compact cluster (37.6–99.1 tokens/step, $n = 10$) from a verbose cluster (282.2–327.0 tokens/step, $n = 3$), with no model between 100 and 282 tokens/step. This gap ($\geq$2.8$\times$ between cluster edges) is specific to GSM8K (Table 27), though GPQA reveals extreme verbosity for GLM-4.7-355B (1,185.7 tokens/step) and Kimi2.5-1T (1,121.4 tokens/step).

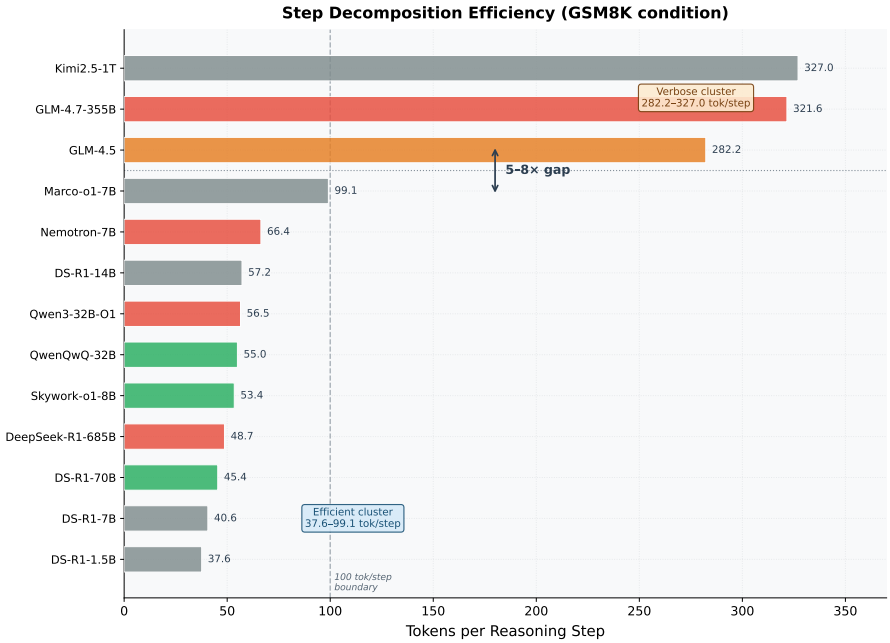



Figure 19: **Step decomposition efficiency across 13 reasoning models (GSM8K condition).** A structural gap ($\geq$2.8$\times$) separates a compact cluster (37.6–99.1 tokens/step, $n = 10$) from a verbose cluster (282.2–327.0 tokens/step, $n = 3$), with no model between 100 and 282 tokens/step. Cross-dataset analysis in Table 27.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All four claims are supported by statistical tests with exact $P$-values: (1) 1,000-token convergence in 21/24 conditions (Section 3.1), (2) +15.4pp causal gain ($P = 0.010$, Section 3.1), (3) 18.8$\times$ efficiency ratio at matched 7–8B scales (Section 3.2), (4) Q-score correlation $r = +0.787$, $P = 0.0014$ (Section 3.2). Scope limitations stated in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 dedicates a full paragraph to limitations: Q-score validated on only 4 held-out models (95% CI [39.8%, 100%]), GSM8K-specific calibration with cross-dataset shifts up to 0.74 points, observational mechanistic dimensions, underpowered McNemar tests (3/21 significant), $n = 50$ per condition limiting detection to effects $\geq$18pp, restriction to open-source models, and limited error annotation scope ($n = 58$, $\kappa = 0.886$, two categories).tificationTODO

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical. No theorems or formal proofs are presented. All claims are supported by statistical tests (paired $t$-tests, Pearson/Spearman correlations, Fisher exact tests, McNemar tests) rather than theoretical derivations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2 specifies all 44 models by name and parameter count, 6 datasets with sample sizes ($n = 50$–300), exact prompt templates (Appendix Table), generation parameters (`do_sample=False`, FP16), token limits (300–8,000), and all statistical tests. Appendix provides complete results for all 288+ experimental conditions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymous code repository has been made available as supplementary material containing the complete evaluation framework, Q-score pipeline, analysis scripts, prompt templates, and processed result CSVs. Raw model responses will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Q-score framework: training/test split (9/4 models) in Section 2.5, weighting hyperparameters (0.7/0.3, 0.5/0.5) with sensitivity analysis across 5 weightings (Appendix E), tier thresholds in Section 2.5. Evaluation: all generation parameters in Section 2.1 (`do_sample=False`, FP16, max token limits). Random seed 42 for random-sample experiment (Appendix D.6).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All main claims report exact $P$-values and test statistics: causal intervention ($t_{51} = 2.68$, $P = 0.010$), Q-score correlation ($r = +0.787$, $P = 0.0014$, 95% CI $[0.417, 0.933]$), sign test ($P = 0.008$), gap reduction ($P = 9.3 \times 10^{-12}$). Leave-one-out stability analysis (min $r = 0.749$, max $r = 0.818$). Bonferroni correction across 236 pairwise tests ($\alpha' = 0.00021$). Wilson confidence intervals in Appendix. Power analysis acknowledges $n = 50$ limits detection to effects $\geq$18pp (Section 4).

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All local models evaluated using HuggingFace Transformers in FP16 on NVIDIA GPUs (Section 2.1). API-based models (Gemini-2.0-Flash) evaluated via google free tier API. Estimated total compute: $\sim$92 GPU-hours across three phases. Analysis from cached data requires $<$5 minutes on any machine.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study evaluates publicly available open-source language models on established benchmarks. No human subjects were involved beyond two annotators performing blinded error categorisation. No private or sensitive data was collected or used. All models used under their respective open-source licences.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive: the framework enables reduced inference costs by identifying early-plateau models needing only 1,000 tokens, reducing energy consumption. Negative: the finding that training quality is the primary efficiency determinant could widen the gap between well-resourced organisations and smaller practitioners. Both discussed in Section 4.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets. All 44 evaluated models are publicly available. The Q-score diagnostic framework is a measurement tool, not a generative system, and poses no misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All 44 models cited with original publications. All 6 benchmarks (GSM8K, BoolQ, ARC-Easy, CommonsenseQA, MMLU-STEM, GPQA) cited with original papers. All models used under their respective open-source licences (Apache 2.0, MIT, or model-specific).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The Q-score framework code and processed evaluation data has been released as an anonymous supplementray with README, model configurations, prompt templates, reproduction scripts (`make analysis`), and a verification test suite that confirms all headline numbers.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing was performed. Error annotation ($n = 58$) was conducted by two persons using the blinded protocol described in Appendix D.5, with complete annotation guidelines provided.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted. The study analyses publicly available language model outputs on established benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We use LLMs to fix grammar mistakes and polish paper writeups

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.