Detect, Restart, Recover: Budget-Sensitive Inference for Stuck Mathematical Reasoning

Anonymous ACL submission

Abstract

Under tight generation caps, large reasoning models often spend the entire budget without emitting an extractable answer, even when the trace contains useful partial work. We identify these *finalization failures* as a separable, recoverable failure mode, distinct from wrong-answer reasoning errors, and give a cheap way to recover them. We introduce **Detect-Restart-Gate (DRG)**, a training-free primitive that detects a stuck trace, restarts it with its own text as *previous-attempt context*, and accepts the retry through a conservative pathology gate. A matched fresh-retry baseline and a 2×2 factorial on three models isolate the active ingredient: previous-attempt context, not a retry instruction or an independent sample. Continuing a stuck trace hurts, but restarting with it as reflection helps. On MATH-500, DRG recovers no-answer failures at 0.69–0.81× the generated tokens of self-consistency (SC-2) with $\geq 97\%$ correctness preservation; in deployment, the trigger reduces to a one-line no-answer rule, while the active mechanism remains restart-with-context. A single operating rule, no-answer pressure, tracks the accuracy effect: a strict Pareto improvement on high-pressure Qwen3-8B/14B (+6.9/+7.7pp at lower token cost, McNemar $p < 10^{-6}$), shading to a token-saving operating point at neutral-to-small accuracy cost as pressure falls. Finalization recovery is a measurable, routable failure mode and a practical target for budget-sensitive inference.

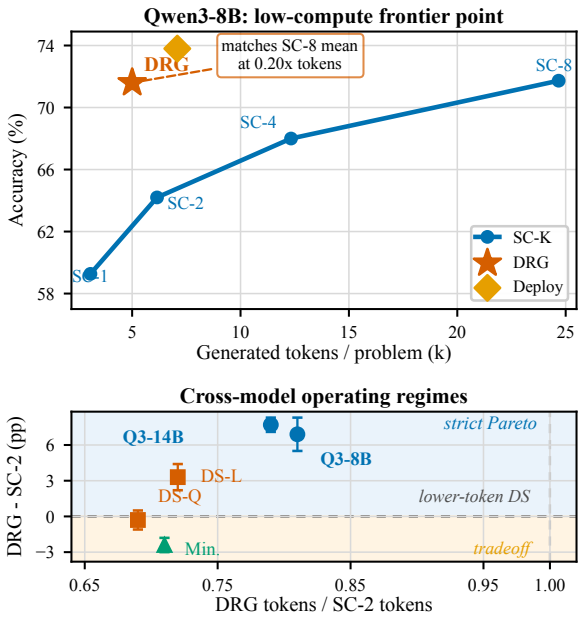

Figure 1: **MATH-500 compute frontier at a 4k cap.** *Top:* on Qwen3-8B, DRG dominates the realized SC-2/SC-4 points and matches the SC-8 mean (3 SC-8 pools) at 0.20× SC-8 tokens; Deploy denotes DRG plus an SC-2 fallback on residual no-answer outputs. *Bottom:* regime-dependent across five models — strict Pareto on Qwen3, lower-token operating point on DS-R1 distills, accuracy–compute tradeoff on Ministral. Error bars: 5-seed DRG retry variation (Qwen3-8B SC-2 also 5-seed).

1 Introduction

Large reasoning models (LRMs) such as DeepSeek-R1 (DeepSeek-AI, 2025), Qwen3 (Yang et al., 2025), and Ministral-3-14B-Reasoning can spend thousands of tokens on multi-step reasoning (Wei et al., 2022; Kojima et al., 2022). Yet long reasoning is only useful if it reaches a final answer. Under tight token caps, LRMs often fail by continuing an unproductive trajectory: repeating calculations, restating assumptions, or cycling through equivalent approaches until generation stops. On Qwen3-8B with a 4k per-generation cap, 38% of MATH-500 generations produce no extractable answer; on AIME 2024 under the same cap, every baseline generation fails to finalize. We call a generation a *finalization failure* when, under our primary strict extractor, it yields no extractable final answer before the cap; cap-hit and permissive-extraction (unboxed-answer) audits are reported separately (Appendix A.28), since a strict no-answer may reflect either genuine non-finalization or an extrac-

tion artifact.

The standard response is to spend more test-time compute through additional samples. This is natural when the problem is lack of diversity, but budget failures are not homogeneous. A stuck trace may already contain useful partial work, so a fresh independent sample is not the only possible intervention. The allocation question is therefore not just *how much* compute to spend, but *which failure mode* the next unit of compute should address.

We study this question through **Detect-Restart-Gate (DRG)**, a training-free selective recovery primitive.[1] DRG detects stuck traces from text-level signals, restarts with previous-attempt context, and accepts the retry only through a conservative pathology gate; in deployment this reduces to a one-line rule — retry iff the baseline emits no extractable answer (or hits the cap). This turns budget exhaustion into a routing decision: some traces should be recovered, some left unchanged, and some remaining failures may still require sampling. We focus on the 4k per-generation cap as a controlled budget-pressure slice and proxy for fixed-output-cap serving, where raising the cap is itself a competing allocation choice; high-cap probes show the boundary, where pressure falls and SC-2 can catch up or win.

Across MATH-500 and AIME/AMC probes, the general result is a recovery one: under budget pressure DRG recovers no-answer failures at lower average generated-token cost than SC-2, with near-zero correctness regression. A 2×2 factorial on three models spanning Qwen3 and DS-R1 distills shows that stuck text is harmful as a continuation prefix but useful as reflection context for a fresh attempt; the explicit retry instruction is not what helps. The accuracy consequence follows no-answer pressure, in one line: DRG *helps* when pressure is high (strict Pareto over SC-2 on Qwen3), is *neutral* when the model already finalizes reliably (equal accuracy at fewer tokens), and *trades* accuracy for tokens at low pressure.

We make three contributions:

Finalization recovery is a separable failure mode. Under tight caps, many failures are *finalization* failures — the baseline has useful partial work but emits no extractable answer. We show these are recoverable, and a matched fresh-retry baseline plus a 2×2 factorial on Qwen3 and DS-R1 distills identify *previous-attempt context* — not retry instructions or independent resampling — as the useful information channel in this retry design.

A training-free recovery primitive with a simple deployable trigger. Across five models on MATH-500, DRG recovers no-answer failures at 0.69–$0.81\times$ SC-2 generated tokens with $\geq 97\%$ correctness preservation. An ablation (Table 5) shows the multi-signal detector reduces to a *no-answer / cap-hit* trigger that matches full DRG accuracy at roughly half the retries: the substantive mechanism is restart-with-context, and the trigger is its deployable wrapper.

An operating rule and a low-token frontier point. The accuracy consequence tracks no-answer pressure (per-family decomposition, Appendix A.10): strict Pareto on high-pressure Qwen3-8B/14B (+6.9/+7.7pp, McNemar $p < 10^{-6}$), a smaller win (Llama-8B) or no detected difference (Qwen-7B) on lower-pressure DS-R1 distills, and a tradeoff on low-pressure Ministral where SC-2 wins accuracy while DRG saves tokens. A realized SC-K frontier places DRG at a low-token point uniform sampling does not reach; the saving is a throughput/serving-cost advantage. The optional Deploy variant routes residual no-answers to SC-2 when extra compute is acceptable.

2 Related Work

Reasoning prompting and efficiency. Chain-of-thought prompting and its variants (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023; Yao et al., 2023; Zelikman et al., 2022; Zhou et al., 2024) optimize the *initial* reasoning trajectory. Work on reasoning efficiency / overthinking reduces trace length by fine-tuning (Chen et al., 2025) or compute allocation (Snell et al., 2025; Wu et al., 2025). DRG is downstream of both: it intervenes training-free on an already-running trace when budget exhaustion itself is the failure mode.

Sampling-based test-time compute. Self-consistency (Wang et al., 2023; Chen et al., 2024) draws K independent generations at $K\times$ cost, while best-of-n with verifiers (Lightman et al., 2024; Wang et al., 2024) requires reward training. Broader scaling work studies repeated sampling and adaptive allocation (Brown et al., 2024; Wu et al., 2025; Snell et al., 2025). DRG contributes a verifier-free selective allocation rule keyed to text-level budget-exhaustion pathology rather than paying for independent samples on every problem.

[1] Code and per-run records (anonymized for review): https://github.com/AnonymousAuthor0211/detect-restart-gate.

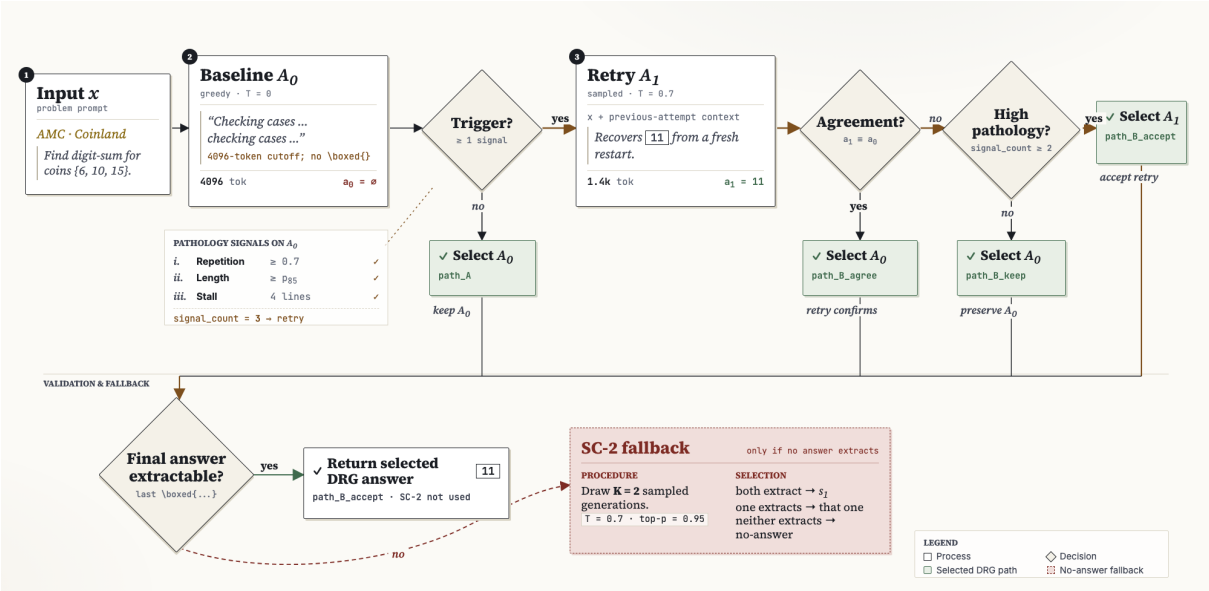

Figure 2: **Detect-Restart-Gate (DRG) and the optional Deploy fallback.** Starting from problem x, DRG detects stuck baseline reasoning A_0, restarts with x plus previous-attempt context to produce A_1, then returns A_0 or A_1 through agreement and pathology gates. The bottom extraction stage invokes SC-2 only when the selected DRG output has no extractable answer; Deploy does not route between disagreeing formatted answers.

We therefore compare primarily against SC-K and a matched adaptive fresh-retry baseline; verifier-trained and critique/tool-based routers are complementary rather than direct zero-training baselines.

Self-correction, reflection, and trace pathology. Self-refinement (Madaan et al., 2023; Huang et al., 2023), Reflexion (Shinn et al., 2023), and critique/tool methods (Miao et al., 2024; Gou et al., 2024) add external signals; Huang et al. (2024) argue that intrinsic self-correction is unreliable. Our factorial analysis refines this in the studied math setting: continuing a stuck trace is harmful, while a fresh generation can benefit from prior-attempt context. Degenerate repetition is a known text-generation failure (Holtzman et al., 2020; Fu et al., 2021); WordSaladChopper trains a classifier to chop repetitive segments for length efficiency (Xie et al., 2025). DRG instead uses an untrained trigger, a fresh retry with prior-attempt context, and a correctness gate.

3 Method: Detect-Restart-Gate

3.1 Overview

Given a problem x, model M, and token budget B, DRG produces an answer in three stages (Figure 2): **detection** of stuck reasoning from baseline A_0, **retry** A_1 conditioned on previous-attempt context, and **pathology-gated agreement** that accepts A_1 only when answers disagree and pathology evidence is strong. The Deploy variant adds a no-answer fallback to SC-2 (Section 4.2). Although full DRG computes three stuckness signals, the deployable trigger is simply no-answer/cap-hit; the richer detector is retained as a conservative diagnostic and for the gate/pathology analysis (Table 5). Full pseudocode is provided in Appendix A.1.

3.2 Detection, Retry, and Gating

Detection. Three text-level signals are computed on the baseline A_0: (i) the whitespace-delimited unigram repetition rate $\mathrm{rep}(A_0) = 1 - |\mathrm{unique_unigrams}(A_0)|/|\mathrm{all_unigrams}(A_0)|$; (ii) the length percentile $\mathrm{len}(A_0)$ of A_0's token count within the per-run baseline distribution; and (iii) a terminal stall flag $\mathrm{stall}(A_0, k)$ that fires when the final nonempty lines stop introducing new lexical terms. Let $s(A_0)$ denote the three thresholded signal indicators:

$$\begin{aligned} s_{\mathrm{rep}} &= \mathbf{1}[\mathrm{rep}(A_0) \geq \tau_{\mathrm{rep}}], \\ s_{\mathrm{len}} &= \mathbf{1}[\mathrm{len}(A_0) \geq P_{\mathrm{len}}], \\ s_{\mathrm{stall}} &= \mathbf{1}[\mathrm{stall}(A_0, k)], \\ \mathrm{trigger}(s) &\Leftrightarrow s_{\mathrm{rep}} \vee s_{\mathrm{len}} \vee s_{\mathrm{stall}}. \end{aligned} \tag{1}$$

Thresholds τ_{rep}=0.7, P_{len}=85th, and k=4 are fixed unlabeled heuristics rather than label-optimized settings: high duplicate unigram mass, top-tail length, and terminal novelty stall. They were fixed prior to benchmark evaluation and reused identically across

Model	B0 acc.	SC-2	DRG	Deploy	Δ	Tok. D/S	Ratio D/Dep	No-Ans
Qwen3-8B	59.6%	64.7±1.3%	71.6±0.4%	**73.8±0.5%**	**+6.9±1.4pp**	5001/6159	0.81/1.15×	38.2%
Qwen3-14B	64.6%	67.6%	75.3±0.6%	**76.8±0.4%**	**+7.7±0.6pp**	4615/5837	0.79/1.10×	33.6%
DS-R1-Qwen-7B	71.8%	79.4%	79.1±0.8%	**82.5±0.4%**	−0.3±0.8pp	3341/4813	0.69/0.94×	22.8%
DS-R1-Llama-8B	70.6%	74.4±1.1%	77.7±0.2%	**79.6±0.5%**	+3.3±1.1pp	3298/4576	0.72/0.92×	21.0%
Ministral-14B	84.8%	90.2%	87.8±0.6%	90.0±0.5%	−2.4±0.6pp	2241/3144	0.71/0.84×	9.2%

Table 1: MATH-500 at a 4k per-generation cap. DRG/Deploy/Δ are 5-seed retry means; B0 is deterministic. SC-2 seed coverage varies by cell, with audits summarized in Appendix A.5. Tok. D/S reports average generated tokens for DRG vs. SC-2; Ratio D/Dep reports DRG and Deploy token ratios relative to SC-2.

every (model, dataset, budget) cell; no threshold uses labels, gold answers, or correctness outcomes. The length percentile is the only distributional parameter and uses unlabeled B0 token counts; calibration details are in Appendix A.12. Exact tokenization for the stall signal is provided with the other implementation details in Appendix A.4. We treat the detector primarily as a compute-routing mechanism, not an accuracy oracle: random triggers underperform real triggers, universal triggering can match selective accuracy on Qwen3 only by spending much more retry compute, and signal ablations show that repetition/length combinations are robust while stall-only is intentionally conservative (Appendix A.8).

Retry. On trigger, the retry prompt contains the original problem and the final 1200 characters of A_0 as previous-attempt context, generated at decoding temperature τ_{dec}=0.7 and top_p=0.95 (the symbol τ_{dec} is used for sampling temperature to keep B for the token budget). This window was fixed before benchmark evaluation and used for all headline results. A post-hoc 3-seed sweep on Qwen3-8B (0, 300, 600, 1200, 2400 chars; Appendix A.11) shows that nonzero windows outperform the zero-context endpoint and that 2400 characters improves further; we treat that row as optimization headroom requiring held-out selection rather than changing the headline setting after observing test performance. No restart instruction is added — Section 4.3 shows explicit instructions are neutral or harmful in this scaffold.

Pathology-gated agreement. Let $\text{path}(s) \Leftrightarrow \text{signal_count}(s) \geq \tau_{\text{path}}$ count how many of the three trigger signals fire on A_0 (the signal-count threshold τ_{path} is distinct from the repetition threshold τ_{rep} of Eq. 1 and the decoding temperature τ_{dec}). Let a_0, a_1 denote the extracted answers of A_0, A_1, and let $\text{ext}(a_i)$ indicate that extraction returned a nonempty answer. Symbolic agreement is defined only when both answers are extractable: $g \Leftrightarrow \text{ext}(a_0) \wedge \text{ext}(a_1) \wedge (a_0 \equiv a_1)$, where $\equiv$ is SymPy equivalence. The accepted output is

$$\text{gate}(A_0, A_1, s) = \begin{cases} A_0 & g, \\ A_1 & \neg g \wedge \text{path}(s), \\ A_0 & \text{otherwise.} \end{cases} \quad (2)$$

Here $\text{gate}(\cdot)$ returns the accepted *trace* $\hat{A} \in \{A_0, A_1\}$; the reported answer $\hat{a}$ is the extracted answer of $\hat{A}$ (under Deploy, the optional SC-2 fallback is invoked only when $\hat{a}$ is empty). If exactly one answer is extractable, or neither is extractable, the pair is treated as non-agreement; a high-pathology baseline therefore selects A_1, while a low-pathology baseline keeps A_0. The agreement branch never overrides a baseline that converged with the retry; the pathology branch overrides only on non-agreement and only when at least τ_{path}=2 signals fire on A_0. This rule achieves near-zero correctness regression: per-model C→W flips are ≤ 4 with $\geq 97\%$ preservation across all five models (Table 27). Sensitivity to τ_{path} and to trigger-signal combinations is in Appendix A.8.

Deployable no-answer fallback. A deployable variant DRG→SC-2 routes only DRG outputs with no extractable answer to an SC-2 fallback. The fallback is deliberately conservative: it captures the no-answer slice of residual failures but does not choose between disagreeing formatted answers.

4 Experiments

4.1 Setup

Models and data. We evaluate five models from three families: Qwen3-8B/14B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-7B/Llama-8B (DeepSeek-AI, 2025), and Ministral 3 14B (served id `ministral-14b-2512`; HF checkpoint `mistralai/Ministral-3-14B-Reasoning-2512`).[2] The primary benchmark is MATH-500 (Hendrycks et al., 2021); AIME 2024 (`datasets/aime_2024`,

[2]Mistral Open release v25.12 (Dec 2025); reasoning-tuned 14B.

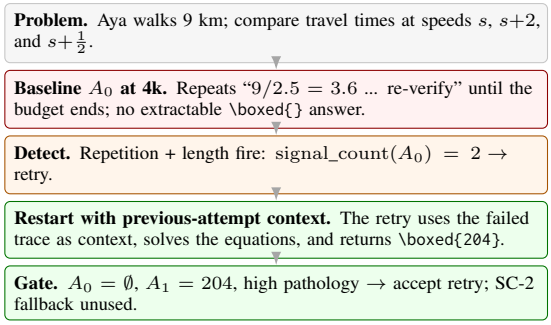

Figure 3: **A compact DRG rescue.** A budget-exhausted baseline contains enough partial structure to guide a restarted attempt, while the pathology gate accepts the retry only when the original trace is both disagreeing/no-answer and visibly stuck.

$n = 30$) and a 40-problem AMC probe (`datasets/amc`; exact problem list and source ids in the artifact) provide budget-pressure probes, with GSM8K, GPQA Diamond, and OlympiadBench Physics (He et al., 2024) as boundary checks. Trigger thresholds, gates, and retry settings are fixed across models.

Generation and baselines. We write B0 for the baseline greedy-generation method; table columns labeled B0 acc. report its accuracy, and B0 no-answer rate is the pressure proxy. MATH-500 and 4k probes use `max_new_tokens=4096` per generation; high-budget probes use 32768. DRG retries use $\tau_{\text{dec}} = 0.7$, top_p $= 0.95$. We compare to SC-2/SC-8 self-consistency at the same sampling settings and to an adaptive fresh-retry baseline that shares DRG's detector, retry budget, and gate but removes previous-attempt context.

Accounting and statistics. We report average generated tokens per problem: SC-K always draws K samples, while DRG retries only triggered cases and Deploy adds SC-2 only on residual no-answer outputs. Input-token accounting changes DRG/SC-2 ratios by at most 3pp (Appendix A.4). Final answers use last-boxed extraction and symbolic equivalence checking; full prompt, extraction, scoring, and seed details are in Appendix A.4. DRG/Deploy/Oracle are 5-seed means for all MATH-500 models and 4k AIME/AMC Qwen3/Ministral cells. SC-2 seed coverage and audits are summarized in Appendix A.5; high-budget AIME/AMC cells are single-seed diagnostics.

Model	Fresh	SC-2	DRG	Tok.	Ctx
Qwen3-8B	64.8%	63.8%	71.6±0.4%	0.90×	+6.8pp
Qwen3-14B	68.4%	67.6%	75.3±0.6%	0.88×	+6.9pp
Ministral-14B	86.4%	90.2%	87.8±0.6%	0.78×	+1.4pp

Table 2: **Budget-matched adaptive fresh-retry baseline.** Fresh uses the same detector, retry budget, and gate as DRG, but retries from the original problem only; it is the zero-context endpoint against the default 1200-character context window. **Tok.** is Fresh generated tokens / DRG generated tokens; **Ctx** is DRG accuracy − Fresh accuracy. Fresh and SC-2 are matched diagnostic comparators; full DRG is the headline mean from Table 1.

Retry Condition	Q3-8B	Q7B	L8B
Clean retry†	64.2%	77.2%	75.4%
Instruction	63.6%	76.6%	73.2%
Prev. context	**71.8%**	**79.8%**	77.2%
Full scaffold	**71.8%**	79.4%	**78.8%**

Table 3: Factorial decomposition of the retry scaffold on MATH-500. Q3=Qwen3, Q7B=DS-R1-Qwen-7B, L8B=DS-R1-Llama-8B. Previous-attempt context drives the gain across these three models spanning Qwen3 and DS-R1 distills; explicit retry instructions are neutral or harmful. †Clean retry is the matched adaptive baseline with the same detector, retry budget, and gate, but no previous-attempt context. This is a matched diagnostic run; Table 2 is a separate matched diagnostic, so the overlapping Q3 fresh/clean value differs slightly.

4.2 MATH-500: DRG and the Deployable Pipeline

Headline (5-seed MATH-500): DRG defines a lower-token recovery operating point on budget-pressured MATH-500. Whether this becomes strict Pareto, no detected difference, or a compute–accuracy tradeoff tracks the per-family pathology rate. On MATH-500 (Table 1), DRG uses 0.69–0.81× SC-2 generated tokens across all five models and three families. **(1) Qwen3 (strict Pareto):** on Qwen3-8B/14B (no-answer 33–38%), DRG significantly beats SC-2 (+6.9±1.4pp / +7.7±0.6pp; McNemar $p < 10^{-6}$ every seed; bootstrap 95% CIs $[+5.0, +10.2]$pp and $[+4.8, +10.4]$pp) — higher accuracy *and* lower compute. **(2) DS-R1 distills (lower-token, mixed accuracy):** both distills run at $\sim$ 0.7× SC-2 tokens. On Qwen-7B, DRG is a lower-token comparable-accuracy point (−0.3±0.8pp; bootstrap CI brackets zero). On Llama-8B, a full 5-seed SC-2 audit places DRG +3.3±1.1pp above SC-2 at 0.72× tokens. **(3) Ministral (compute–accuracy tradeoff):** at 9% no-answer (lowest pressure), standalone DRG

Bench.	Model	Bud.	B0 acc.	DRG	SC-2	Deploy	Oracle	Regime
AIME	Qwen3-8B	4k	0.0	**18.0±3.8%**	3.3	**18.0±3.8%**	18.0	R
AIME	Qwen3-14B	4k	6.7	**15.3±3.0%**	6.7	**18.7±3.0%**	18.7±3.0	R
AIME	Ministral-14B	4k	36.7	42.0±1.8%	43.3	**50.7±1.5%**	51.3±1.8	C^*
AMC	Qwen3-8B	4k	30.0	**42.5±1.8%**	35.0	**45.0±1.8%**	45.0	R
AMC	Qwen3-14B	4k	40.0	**47.5±3.1%**	35.0	**47.5±3.1%**	47.5±3.1	R
AMC	Ministral-14B	4k	65.0	**82.0±3.3%**	77.5	**84.5±2.1%**	84.5±2.1	R

Table 4: Cross-benchmark AIME/AMC accuracy (%) at a 4k cap; 5-seed DRG/Deploy/Oracle means for Qwen3 and Ministral. These 4k cells probe whether the recovery pattern appears under high pressure; the recovery regime (R) is the dominant label (C^* = a close standalone cell rescued by Deploy). The single-seed 32k diagnostics and the full exploratory regime taxonomy (R/C/V/X) are deferred to Appendix A.19 (Table 23).

Model	B0 acc.	DRG	No-ans only	Cap hit	Trig% D/NA	Tok× D/NA
Qwen3-8B†	59.6	71.8	**72.2**	71.8	72/38	.81/.72
Qwen3-14B	64.6	75.3	**75.3**	75.3	63/34	.79/.71
DS-Qwen-7B	71.8	79.1	**79.4**	79.1	50/22	.69/.60
DS-Llama-8B	70.6	77.7	**78.2**	77.7	53/21	.72/.62
Ministral-14B	84.8	87.8	**88.1**	88.1	32/9	.71/.60

Table 5: **The deployable trigger simplifies to no-answer / cap-hit.** Restricting DRG's *trigger* to no-answer (or near-identically cap-hit) baselines — with the restart and the $\geq$ 2-signal pathology gate unchanged from full DRG (Algorithm 1) — reproduces full-DRG accuracy at roughly half the trigger rate and lower token cost. Accuracy is 5-seed mean (†Qwen3-8B is 4-seed, seeds 0–3; reproduces Table 1 within noise); no-answer-only vs. full DRG is McNemar n.s. on all five models ($p > 0.1$). **Trig%**: trigger rate, full DRG / no-answer-only. **Tok×**: generated tokens relative to SC-2. Correctness regression (C→W) is ≤ 3 per model.

trades −2.4±0.6pp accuracy for 29% fewer tokens (per-seed $p \in [0.0037, 0.108]$); Deploy recovers this cell to 90.0±0.5% vs. 90.2% for SC-2 at 0.84× SC-2 tokens.

The deployable trigger is no-answer / cap-hit. In the audited fix-type runs (Qwen3-8B, DS-R1-Qwen-7B), every wrong→correct fix is a finalization recovery on a no-answer baseline, not a wrong-answer correction (Appendix A.14). Restricting the trigger to no-answer baselines therefore reproduces full DRG accuracy — McNemar not significant on all five models — at roughly half the retries and 0.60–0.72× SC-2 tokens (Table 5); cap-hit triggering is near-identical. The repetition/length/stall signals are a conservative diagnostic superset, but in deployment the rule simplifies to "retry the previous attempt iff it produced no extractable answer." We report this saving as generated tokens, the relevant cost in throughput-bound serving; parallel-serving latency is analyzed in Appendix A.13. A cheaper forced-finalization control — inject an end-of-thinking + answer cue onto the stuck prefix and generate $\leq$ 256 tokens — fixes only 25/7/0 no-answer cases on Qwen3-8B/14B/DS-R1-Qwen-7B versus DRG's 63/52/38: asking the model to finalize rarely suffices, while a fresh restart with previous-attempt context recovers many more cases (Appendix A.15).

Two operating modes: DRG (lower compute) and Deploy (accuracy fallback). Standalone DRG is the lower-average-token primitive; Deploy adds an SC-2 fallback only on residual no-answer outputs. Deploy improves accuracy on the no-answer slice at 0.84–1.15× SC-2 tokens, with per-model rescue counts and the Pareto plot in Appendix A.21.

Matched fresh-retry baseline identifies the role of context. A clean adaptive baseline — same detector, same retry budget, but no previous-attempt context — gives Qwen3-8B / Qwen3-14B / Ministral-14B accuracies of 64.8% / 68.4% / 86.4%, only +1.0/+0.8/−3.8pp over SC-2 (Table 2). The matched clean-retry baseline also restarts from the original problem and can equally repair generic finalization or answer-format failures; the additional +6.8/+6.9pp from adding previous-attempt context on Qwen3 therefore cannot be explained by restart or answer-format repair alone. The factorial (Table 3) shows the same directional pattern on the DS-R1 distills, with smaller gains under lower no-answer pressure. A 3-seed window-length sensitivity check on Qwen3-8B confirms the context effect: all nonzero windows outperform the zero-context endpoint while C→W flips remain rare (Appendix A.11).

Diagnostic decomposition — why do gains differ across families? Decomposing the DRG–B0 gain as (trigger rate) × (per-trigger fix rate), the per-trigger fix rate is stable (12.6–17.3%) across three families while trigger rate varies 2× (31.8% Ministral → 71.6% Qwen3-8B; Table 13, Appendix A.10). The gain thus tracks recoverable-pathology frequency: regressing the DRG–SC-2

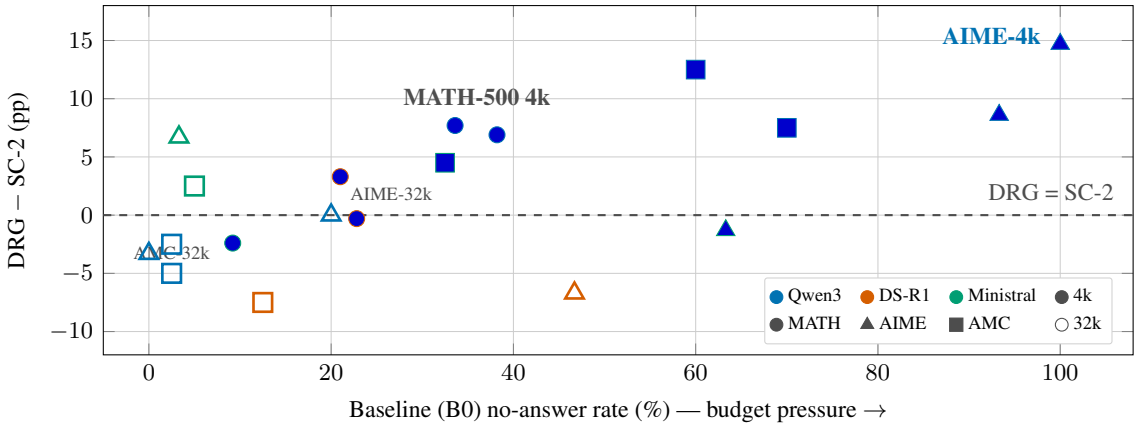

Figure 4: **Operating regimes (DRG vs. SC-2).** Each point = (model, benchmark, cap); x = B0 no-answer rate, y = DRG−SC-2 (pp). Filled 4k cells (5-seed for Qwen3/Ministral) show the pressure pattern: high pressure tends to recover ($y > 0$), lower pressure moves toward neutral/tradeoff. Open 32k single-seed points are visual boundary diagnostics; the full taxonomy is exploratory.

gain on B0 no-answer rate across the five 4k cells (strict extraction) gives slope $+0.35$pp/pp (Spearman 0.80), crossing zero near 16% no-answer. This supports pressure as the main operating signal, while family and sampling variance still matter; Ministral's loss (9.2% no-answer, -2.4pp) sits on the low-pressure side of this diagnostic (Appendix A.28).

Frontier audit — realized SC-K frontier: DRG occupies a useful low-compute frontier point. Against multi-pool SC-K frontiers (3 SC-8 pools; Appendix A.16), DRG matches or exceeds the SC-8 *mean* at $0.20\times$ SC-8 compute on *both* Qwen3 models (SC-8 std ≤ 0.4pp) and dominates SC-2/SC-4; on DS-R1-Qwen-7B it dominates the SC-2 point while SC-4/8 buy higher accuracy at substantially higher token cost. Uniform sampling reaches comparable accuracy only at much higher K/token cost, and at $K \in \{1, 2, 4\}$ does not reach DRG's low-token recovery point. DRG is thus a low-token recovery operating point, not a universal accuracy dominator — larger pools or caps match or exceed its accuracy at substantially higher token cost. A uniform permissive-extraction audit on Qwen3-8B shows the advantage over SC-2 persists ($+5.7$pp, while all methods' absolute accuracies rise; Appendix A.28).

4.3 Why DRG Works: Restart plus Previous-Attempt Context

Why does detection-based retry produce a different solution set than independent sampling? A retry can add two ingredients beyond the original problem: (1) an explicit restart instruction and (2) the final 1200 characters of A_0 as previous-attempt context. The default DRG condition uses the second ingredient only; the fresh-retry baseline tests the 0-character endpoint, and the factorial decomposes instruction vs. context on three models spanning Qwen3 and DS-R1 distills (MATH-500, all 500 problems, identical trigger and gate settings).

Three findings hold across these three factorial models: **(1) Instruction alone is counterproductive** (-0.6pp Qwen3-8B, -0.6pp Qwen-7B, -2.2pp Llama-8B). Explicit "try again" instructions do not help and may prime defensive or repetitive reasoning, consistent with the 8-operator equivalence analysis (Appendix A.7), where negative-feedback phrasings are the only operator class worse than a neutral restart (-5.6pp, $p = 0.014$, not surviving Bonferroni at 0.00625; suggestive rather than conclusive). **(2) Previous-attempt context is the dominant beneficial factor in this matched retry design**, scaling with stuck severity: +7.6pp Qwen3-8B (38% no-answer), +2.6pp Qwen-7B (23%), +1.8pp Llama-8B. **(3) The instruction–context interaction is secondary**: full scaffold matches previous-only on Qwen3-8B, is slightly worse on Qwen-7B, slightly better on Llama-8B.

Continuation–reflection pattern, scoped. A perturbation ladder (Appendix A.25) shows that *continuing from* stuck text traps the model: accuracy improves as stuck continuation text is removed, and full restart adds formatting reliability (133/135 ties on the matched-format subset). This establishes the first half of the pattern: not continuing the corrupted trace is itself valuable. The matched fresh-retry baseline and factorial estab-

lish the second half: among fresh restarts with the same detector, retry budget, and gate, previous-attempt context adds accuracy beyond a no-context retry. We therefore use "continuation–reflection" descriptively for this math-reasoning setting, not as a universal LRM mechanism: recovery reuses information from a failed attempt, while sampling explores independent starting points.

4.4 Cross-Benchmark Operating Regimes

We evaluate DRG, SC-2, and Deploy on AIME 2024 ($n = 30$) and AMC ($n = 40$) at 4k/32k caps across Qwen3 and Ministral, plus DS-R1-Qwen-7B at 32k. Figure 4 plots all 19 settings; Table 4 lists the multi-seed 4k pressure probes, with 32k single-seed diagnostics tabulated in Appendix A.19.

The main claim is the pressure rule in the multi-seed cells — DRG helps where no-answer pressure is high — while the R/C/V/X taxonomy remains an *exploratory* boundary illustration (Appendix A.19).

Cross-benchmark pressure probes. AMC-4k reproduces the recovery cluster on Qwen3 (+7.5±1.8pp, +12.5±3.1pp) and Ministral-14B (+4.5±3.3pp standalone; Deploy 84.5±2.1%). On Ministral AIME-4k (63.3% no-answer), standalone DRG and SC-2 finish within ±2pp while Deploy adds no-answer recoveries (50.7±1.5%, within reported seed variation of the oracle). Extreme-pressure Qwen3 AIME-4k cells (100% / 93.3% no-answer; B0 0% / 6.7%) stress-test near-total non-finalization: DRG reaches 18.0±3.8% / 15.3±3.0% (+14.7±3.8pp, +8.6±3.0pp), and Deploy reaches the any-path oracle on Qwen3-14B. Overall, retry beats sampling where budget exhaustion is common, and the standalone gap narrows as pressure falls.

Controls. Trigger ablations, real-vs-random controls, operator tests, forced finalization, permissive extraction, and $\geq 97\%$ correctness preservation support the mechanism rather than a prompt artifact (Appendix A.23). At 32k, single-seed diagnostics illustrate boundary cases where gaps narrow/reverse or residual headroom shifts to disagreement routing (Appendix A.19).

5 Conclusion

Under tight caps, many failures are *finalization* failures: useful partial work without an extractable answer. Matched baselines and a 2×2 factorial show that recovery comes from restarting with failed-attempt context, not an explicit retry instruction or an independent sample. The deployable recipe is small: retry no-answer/cap-hit baselines with their own text as context, then gate. It recovers such failures cheaply under high no-answer pressure; when baselines already finalize, DRG is token-saving or neutral and self-consistency can win accuracy. Finalization recovery is therefore a measurable routing primitive, not a universal replacement for sampling.

Limitations

We evaluate DRG primarily on mathematical reasoning under tight caps; GSM8K, GPQA, OlympiadBench, and high-budget probes mark boundary behavior rather than broad NLP generalization. DRG is most useful at high no-answer pressure; as pressure falls, SC-K can match or exceed accuracy, and below a capability floor there may be little recoverable partial work. Thresholds are fixed unlabeled heuristics; length calibration, context-window sensitivity, SC-2 seed audits, and high-budget diagnostics are detailed in Appendix A.3. We report generated-token cost for throughput/serving-cost settings; latency-optimized parallel SC-K is analyzed in Appendix A.13. Public benchmark contamination cannot be ruled out, so claims concern relative inference routing under fixed models/evaluators, not absolute reasoning ability.

References

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

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. 2025. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9487–9499. PMLR.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *AAAI Conference on Artificial Intelligence*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *International Conference on Learning Representations (ICLR)*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling (COLM)*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations (ICLR)*.

Wenya Xie, Shaochen Zhong, Hoang Anh Duy Le, Zhaozhuo Xu, Jianwen Xie, and Zirui Liu. 2025. Word salad chopper: Reasoning models waste a ton of decoding budget on useless repetitions, self-knowingly. In *Proceedings of the 2025 Conference*

on Empirical Methods in Natural Language Processing, pages 33588–33598, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chengyuan Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. SELF-DISCOVER: Large language models self-compose reasoning structures. In *Advances in Neural Information Processing Systems*.

A Supplementary Material

A.1 Full DRG Pseudocode

Algorithm 1 Detect-Restart-Gate (DRG)

Require: problem x, model M, budget B, thresholds $\tau_{\text{rep}}, P_{\text{len}}, k, \tau_{\text{path}}$
Ensure: selected trace $\hat{A}$ and extracted answer $\hat{a}$

```
A_0 ← M(x, B, greedy)                                   ▷ baseline
a_0 ← extract(A_0)
s ← (rep(A_0)≥τ_rep, len_pct(A_0)≥P_len, stall(A_0, k))
if ¬(s_1 ∨ s_2 ∨ s_3) then return (A_0, a_0)            ▷ no trigger
end if
A_1 ← M(x‖tail_1200(A_0), B; τ_dec=0.7, p=0.95)         ▷ 1200 chars, not tokens
a_1 ← extract(A_1)
agree ← (a_0 ≠ ∅) ∧ (a_1 ≠ ∅) ∧ (a_0 ≡_sympy a_1)
if agree then return (A_0, a_0)                         ▷ extractable agreement
end if
if signal_count(s) ≥ τ_path then return (A_1, a_1)      ▷ pathology override on non-agreement
else return (A_0, a_0)
end if
```

A.2 Deployment Recipe

The deployable form of DRG reduces to four steps (no detector tuning, no training):

1. **Run** the baseline A_0 once (greedy, fixed cap B).

2. **Trigger** iff A_0 has *no extractable answer* (or hit the cap). The richer repetition/length/stall detector is an optional conservative superset; in the evaluated regime it does not change accuracy (Table 5).

3. **Restart**: resample once at τ_{dec}=0.7 from the original problem + the final 1200 characters of A_0 as previous-attempt context (no "try again" instruction).

4. **Gate**: keep A_0 if its answer agrees with the retry; accept the retry only on disagreement when ≥ 2 pathology signals fired; route still-no-answer outputs to an SC-2 fallback if extra compute is acceptable (Deploy).

The no-answer / cap-hit form simplifies *only* Step 2 (the trigger); Steps 3–4 are identical to full DRG, and this is how Table 5 was computed. Thus no-answer baselines still compute the pathology signals for acceptance: with a_0=$\emptyset$, agreement is false and the gate selects the retry exactly when ≥ 2 signals fired; otherwise the empty A_0 remains for optional Deploy fallback. Cap-hit baselines that already finalized are gated exactly as in full DRG.

For a new model/task, estimate B0 no-answer pressure on a small unlabeled validation slice (50–100 representative examples): use standalone DRG in high-pressure settings, prefer SC-2 when no-answer pressure is low, and A/B test DRG versus Deploy in the middle. Expect a generated-token saving in throughput-bound settings and, when the baseline already finalizes reliably, parity rather than an accuracy gain.

A.3 Detailed Limitations

Domain scope. Positive results are demonstrated on mathematical reasoning under budget pressure: MATH-500 (Hendrycks et al., 2021), AIME 2024, and AMC. The text-level trigger signals (repetition, length, stall) are designed for settings where unproductive loops have distinctive textual patterns. A non-math STEM transfer probe (OlympiadBench Physics (He et al., 2024), Appendix A.9) shows the recovery mechanism transfers structurally (W→C>C→W on both Qwen3 models), but absolute gains are small (+1.7–2.5pp) because the model sits near its capability floor on physics olympiad problems at a 4k per-generation cap. On GPQA Diamond (Rein et al., 2024), productive scientific deliberation activates the same signals (61% trigger rate, +1.5pp gain only). Extending the framework with large gains to non-mathematical domains likely requires either stronger models or domain-adapted detection. GSM8K (Cobbe et al., 2021) is the lower boundary: the method mostly abstains (0.86% trigger rate; 13 W→C and 1 C→W, aggregate accuracy unchanged).

Benchmark coverage. AIME 2024 and AMC at both 4k and 32k per-generation caps cover both Qwen3 models (8B and 14B) and Ministral-3-14B; DS-R1-Qwen-7B provides high-cap probes on AIME and AMC at 32k. DS-R1-Llama-8B on AIME/AMC remains unevaluated in the current submission. MATH-500 has full five-model coverage. OlympiadBench Physics ($n = 236$) is evaluated on both Qwen3 models as a non-math STEM transfer probe. Cross-family generalization to the Ministral family is partial under 5-seed evaluation: Ministral-14B replicates the recovery-regime pattern on AMC-4k (+4.5±3.3pp standalone, Deploy 84.5±2.1%) but not on AIME-4k where standalone DRG sits slightly below SC-2 (−1.3±1.8pp) although Deploy still wins (+7.4pp).

Operating-regime dependence. Single-seed 32k diagnostics include cells where standalone DRG sits below SC-2 as no-answer pressure falls.

This is consistent with the operating-regime story — DRG is a recovery method, not a universal accuracy method — but means that practitioners should select between DRG, SC-2, and the Deploy pipeline based on expected budget regime.

Deploy captures only no-extractable-answer outputs. The current deployable pipeline routes only on no-extractable-answer outputs. At high budgets where no-answer rates fall to zero, Deploy degenerates to DRG and the +10pp oracle headroom visible on AIME-32k is not captured. A disagreement-aware router would close this gap but requires a routing signal we do not currently have.

Capability floor. On DeepSeek-R1-Distill-Qwen-1.5B (B0 19.0%), standalone DRG (31.6%) sits below SC-2 (34.2%) because the model lacks enough latent ability to extract recoverable partial work from triggered traces — the recovery regime requires a capability floor. The Deploy variant still gains on this cell (41.0%, $+6.8$pp over SC-2) but at $1.54\times$ tokens: it remains useful when additional compute is acceptable, but is not a free upgrade. Below this capability floor, practitioners should prefer SC-2 or Deploy over standalone DRG.

Thresholds, context length, and seed protocol. We use identical trigger parameters across all models, with no label-based or per-model tuning. These values were selected as simple unlabeled heuristics for high repetition, top-tail length, and terminal novelty stall; we did not run a separate held-out tuning protocol, and model-specific tuning could improve results at the cost of this fixed-threshold setting. The length signal uses a fixed percentile rule but is calibrated from B0 token counts for each run, so deployment requires an online or held-out calibration stream.

The detector should be interpreted primarily as a compute-efficiency filter: universal triggering can match selective accuracy on Qwen3 but uses more retry compute, while random triggering at the same rate underperforms. The 1200-character retry window was fixed before headline evaluation and retained as the pre-specified default. On Qwen3-8B, a 3-seed sweep over $\{0, 300, 600, 1200, 2400\}$ chars shows that all nonzero windows outperform the zero-context endpoint while C→W flips remain rare (Appendix A.11); the 2400-character row is optimization headroom requiring held-out selection before becoming a default.

Retry-seed variance is reported as 5-seed mean ± std for all five models on MATH-500 and for Qwen3-8B/14B and Ministral-14B on AIME/AMC at a 4k per-generation cap (Section 4.1, Tables 1–4). SC-2 is reported as a fixed two-sample comparator with K=2 samples at τ_{dec}=0.7, except on Qwen3-8B and DS-R1-Llama-8B MATH-500 where SC-2 is 5-seed (Qwen3-8B 64.7±1.3%, std 1.3pp; Llama-8B 74.4%, std 1.1pp) and on Qwen3-14B / DS-R1-Qwen-7B where 2-seed audits are reported as robustness checks (68.4±0.8% and 76.5±1.3%; Appendix A.5). High-budget 32k AIME/AMC cells are single-seed diagnostic probes, so we use them to illustrate regimes rather than to make fine-grained claims about small margins.

Statistical scope clarification. The $p = 0.020$ figures reported for Control 2 in Table 28 refer to real vs. random *trigger* selection on three DS-R1/Mistral models with 50 random seeds; Qwen3 family controls (Appendix A.22) use a single random seed and are qualitative.

Risks and data handling. Our text-level trigger could in principle suppress legitimately repetitive reasoning (e.g., iterative algorithms or proof structures with repeated base cases); the gate's near-perfect correctness preservation (Table 27) mitigates but does not eliminate this. We use only public benchmarks (MATH-500, GSM8K, GPQA, AIME 2024, AMC, OlympiadBench Physics) containing no PII or offensive content. Because these benchmarks are public and model training corpora are opaque, benchmark contamination cannot be ruled out; this is why we interpret results as relative inference-routing comparisons under fixed models, not as claims about absolute model capability. In deployment, previous-attempt context should be treated as user-provided data because DRG reuses it in the retry prompt. The method does not generate text beyond what the underlying LRM produces and has no obvious dual-use vector beyond the base model itself.

A.4 Implementation Details for Reproducibility

Prompt construction. All math-style runs use Hugging Face `tokenizer.apply_chat_template(..., add_generation_prompt=True)`. Qwen3 thinking runs pass `enable_thinking=true`; otherwise the tokenizer default is left unchanged. The baseline B0 prompt uses the system message: "You are a helpful math assistant. Think step by step and give the final answer in `\boxed{}`." The user message is the raw problem text. B0 uses greedy decoding (`do_sample=false`).

The default DRG retry is the previous-attempt-only condition. It uses `retry_prompt_mode=user`, `retry_instruction_mode=none`, `retry_previous_attempt_chars=1200`, `retry_previous_attempt_variant=real`, $\tau_{\text{dec}} = 0.7$, and top_p $= 0.95$. The retry user message is exactly:

```
{problem_text}
Previous attempt (may be flawed):
{last_1200_characters_of_baseline_text}
```

No additional instruction is inserted in the default condition. In the restart-instruction control, the inserted instruction is: "Restart from scratch. Try again and give your final answer." In the clean $\tau_{\text{dec}} = 0.7$ control, the previous-attempt window is set to zero and the retry user message is only the problem text.

SC answer selection. SC-2 uses the same prompt as B0 with two independent sampled generations at $\tau_{\text{dec}} = 0.7$, top_p $= 0.95$. If both answers are extractable and equivalent, we return the first sample's answer. If both are extractable and disagree, we tie-break by the first sample. If only one sample is extractable, we return the extractable one; if neither is extractable, the result is no-answer. SC-8 uses the same sampling settings with $K = 8$, clusters answers by mathematical equivalence, returns the largest cluster, and tie-breaks by earliest sample.

Trigger implementation. Unigram repetition is whitespace based:

$$\text{rep}(A_0) = 1 - \frac{|\text{set}(\text{split}_{ws}(A_0))|}{|\text{split}_{ws}(A_0)|}.$$

The length hit is computed from the B0 token-count distribution for the run using the fixed 85th percentile rule. The stall detector strips empty lines, requires at least $k + 1$ nonempty lines, extracts alphabetic terms with `[A-Za-z][A-Za-z0-9_]*`, and fires when none of the last $k = 4$ lines introduces a term not seen in earlier lines. It intentionally does not parse LaTeX or distinguish mathematical variables from English words. The trigger fires if any signal is active; the pathology gate treats the baseline as high-pathology when at least two signals are active.

Extraction and scoring. If a response contains `</think>`, we score only the visible suffix after the final `</think>`. For math-style datasets, the extractor scans this visible text for the last balanced `\boxed{...}` expression and returns its contents; otherwise the prediction is no-answer. GSM8K additionally accepts the last `#### number` pattern when no boxed answer appears. GPQA-style multiple-choice extraction accepts boxed A–D answers or the final textual pattern `answer: A`, `final answer: A`, or `choice A`. Predictions and gold answers are normalized, exact-matched first, then compared with symbolic equivalence under a 5-second timeout. The symbolic equivalence stack falls back through SkyThought `math_equal`, our local `symbolic_math.py`, and SymPy/`latex2sympy2`-based parsing. In the gate (Eq. 2), equivalence is invoked only when both A_0 and A_1 yield extractable answers; if one or both extractions return `None`, agreement is false. A symbolic-equivalence check that times out (5 s) or raises is treated conservatively as non-equivalence ($a_0 \not\equiv a_1$), not as no-answer: `math_answers_equal` wraps the comparison in a timeout and returns `False` on any exception. No-answer arises only when extraction/normalization returns `None` *before* equivalence checking. A timeout therefore routes the gate toward its conservative branches (keep A_0, or accept A_1 only when pathology fires) and never spuriously declares agreement.

Seeds, models, and artifacts. Main reported runs use CLI seed 0; per-generation seeds are deterministic functions of base seed, problem id, sample index, and method tag. For DRG retry the tag is `DRG_RETRY`; for SC-2 the tag is `SC2` with sample indices 0 and 1. DRG, Deploy, and Oracle are reported as 5-seed mean $\pm$ std over retry seeds 1–5 across MATH-500 (all five models) and 4k AIME/AMC cells (Qwen3-8B, Qwen3-14B, Ministral-14B); direction is consistent across all five retry seeds in every reported 5-seed cell. B0 is deterministic. SC-2 is 5-seed on Qwen3-8B (64.7$\pm$1.3%, std 1.3pp) and DS-R1-Llama-8B (74.4%, std 1.1pp) MATH-500, audited with 2 seeds on Qwen3-14B and DS-R1-Qwen-7B (Appendix A.5), and fixed seed 0 otherwise. High-budget AIME/AMC cells are single-seed diagnostics. The random-trigger control uses seed 1729 and samples from a deterministic per-problem RNG.

Checkpoints are official Hugging Face releases: `Qwen/Qwen3-8B`, `Qwen/Qwen3-14B`, `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`, `deepseek-ai/DeepSeek-R1-Distill-Llama-8B`,

and mistralai/Ministral-3-14B-Reasoning-2512. The artifact records the local checkpoint path used by every run, the exact AIME/AMC problem lists, and scripts to checksum local model snapshots; model weights are not redistributed. Primary runs used 4× NVIDIA H100 80GB HBM3; main tables use generated-token accounting, with input-token sensitivity summarized in Section 4.1.

The release bundle includes `run_pipeline.py`, fetch/aggregation scripts, `answer_normalization.py`, `symbolic_math.py`, and per-run records sufficient to recompute triggers, gate decisions, fallback paths, token accounting, McNemar tests, and overlap analyses without rerunning inference. Reviewer-accessible anonymized artifact: an MIT-licensed code repository at https://github.com/AnonymousAuthor0211/detect-restart-gate, to be deanonymized at camera-ready.

A.5 SC-2 Seed Audit

Table 6 reports the completed and partial SC-2 seed audits referenced in Table 1. These audits are not used to claim equivalence in the lower-token DS-R1 cells; they check whether the primary operating-point conclusions depend on a single unlucky SC-2 draw.

Model	Seeds	SC-2 Acc.	Avg Tok.
Qwen3-8B	5	64.7±1.3	6,159±20
DS-R1-Llama-8B	5	74.4±1.1	4,576±31
Qwen3-14B	2	68.4±0.8	5,817±4
DS-R1-Qwen-7B	2	76.5±1.3	4,305±34

Table 6: SC-2 stochastic baseline audit on MATH-500. Qwen3-8B and DS-R1-Llama-8B are complete 5-seed audits used in Table 1; Qwen3-14B and DS-R1-Qwen-7B are 2-seed checks. **Transparency:** the Llama-8B value previously reported from fixed seed 0 (78.6%) sits high in the 5-seed distribution (74.4±1.1%, seeds 75.6/74.8/74.8/72.4/74.6); we therefore report the 5-seed mean, under which DRG (77.7%) exceeds SC-2 at 0.72× tokens. Qwen3-14B DRG remains +6.9pp over the audit mean. For DS-R1-Qwen-7B the 2-seed audit is lower than the fixed-seed comparator in Table 1; we keep the conservative Table 1 comparison and treat this as a lower-token operating-point cell rather than an accuracy-win claim.

A.6 Qualitative DRG Rescue Examples

A.7 Operator Equivalence Details

Full operator-equivalence numbers referenced in Section 4.3: the eight-operator contingency table on the triggered subset, per-operator gate-acceptance breakdown, fix-overlap statistics, Jaccard matrix, and token usage are tabulated in Section A.23 below. The operator strings/descriptions are given in Section A.26, and the Cochran's Q test details in Section A.27.

Operator	Acc	Δ NR	p	W→C/C→W
NR	67.1%	–	–	42 / 1
A1	67.1%	+0.0	0.831	41 / 0
B1	66.7%	−0.4	1.000	41 / 1
Paraphrase	65.1%	−2.0	0.441	37 / 1
WSC-style	64.7%	−2.4	0.239	36 / 1
Step	63.9%	−3.2	0.080	35 / 2
Try	63.5%	−3.6	0.052	34 / 2
Neg.	61.5%	−5.6	**0.014**	30 / 3

Table 7: Operator-equivalence contingency on Qwen-7B MATH-500 triggered subset.

Operator equivalence on Qwen-7B MATH-500 ($n = 252$ triggered).

Operator	Accept	Fix	Fix/Acc.	C→W
NR	108	42	38.9%	1
A1	107	41	38.3%	0
B1	108	41	38.0%	1
WSC-style	108	36	33.3%	1
Paraphrase	107	37	34.6%	1
Step	109	35	32.1%	2
Try	109	34	31.2%	2
Neg.	110	30	27.3%	3

Table 8: Per-operator gate-acceptance breakdown.

Fix rate per gate acceptance. Only 9 problems are fixed by all 8 operators; 18 are fixed by exactly one operator. The union of fixes across all operators is 70, compared to 42 for NR alone. Operators fix somewhat different problems (Jaccard 0.42–0.67), but the gate normalizes outcomes so these differences do not translate to accuracy gains.

A.8 Trigger Signal Ablation Details

Full seven-combination trigger ablation on Qwen-7B MATH-500 (NR + gate) referenced in Section 4.3:

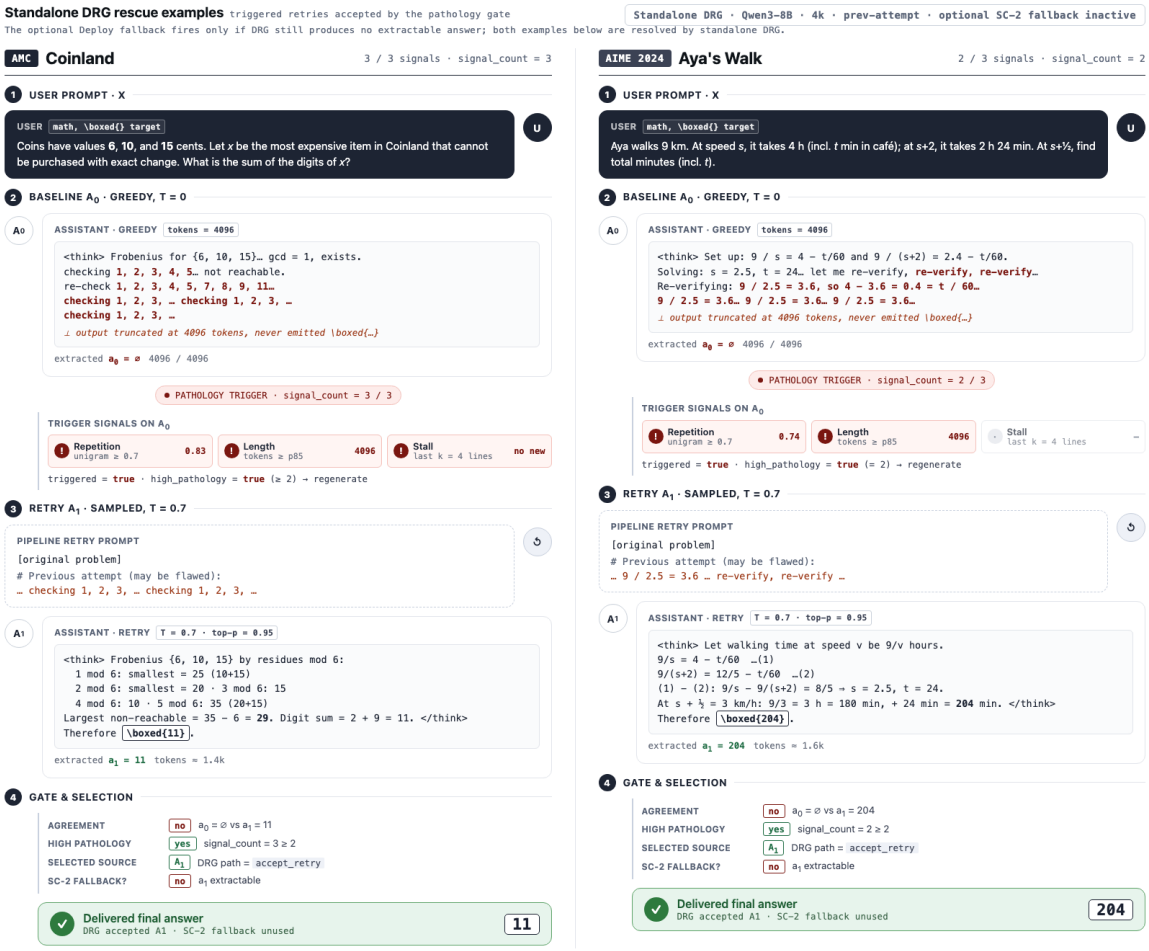

Figure 5: **Illustrative standalone DRG rescue examples.** Two triggered no-answer baselines are restarted with the original problem plus previous-attempt context. In both cases, the retry produces an extractable answer, the pathology gate accepts A_1, and the optional SC-2 fallback is unused.

Signals	Trig.	Acc	Δ B0	Fix	C→W
rep + length	50.4%	**80.8%**	+9.0	47	2
rep + length + stall	50.4%	**80.8%**	+9.0	47	2
length only	25.0%	80.6%	+8.8	45	1
length + stall	25.0%	80.6%	+8.8	45	1
rep + stall	49.2%	80.2%	+8.4	44	2
rep only	48.4%	79.8%	+8.0	42	2
stall only	10.8%	75.6%	+3.8	19	0

Table 9: Trigger signal ablation, Qwen-7B MATH-500.

Metric	Value
Baseline accuracy	84.8%
DRG accuracy	84.8%
Trigger rate	0.86% (76 / 8,792)
Fixes (W→C)	13
Worsens (C→W)	1
Universal-triggering accuracy	81.8% (−3.0pp)

Table 10: GSM8K boundary behavior.

A.9 Boundary Condition Details

Full GSM8K and GPQA Diamond boundary breakdowns:

GSM8K (Qwen-7B).

GPQA Diamond (Qwen-7B, 32k per-generation cap).

Metric	Value
Baseline accuracy	43.4%
DRG accuracy	44.9%
Trigger rate	61.1%
Fixes (W→C)	3
Worsens (C→W)	0
Gate preservation	100%

Table 11: GPQA Diamond boundary behavior.

Metric	Qwen3-8B	Qwen3-14B
N problems	236	236
Baseline accuracy	3.8%	5.1%
DRG accuracy	5.5%	7.2%
SC-2 accuracy	4.2%	5.1%
Deploy accuracy	5.9%	7.6%
Oracle (any-path)	6.4%	8.1%
Trigger rate	97.5%	94.9%
B0 no-answer rate	86.0%	77.5%
DRG no-answer rate	66.5%	56.8%
Fixes (W→C)	5	5
Worsens (C→W)	1	0
Overlap (B/D/SC/N)	8/5/2/221	10/7/2/217

Table 12: OlympiadBench Physics ($n = 236$) boundary breakdown. The recovery mechanism transfers structurally (W→C>C→W on both models, monotonic improvement from 8B to 14B), but absolute gains are bounded by the low baseline capability on physics olympiad problems at the 4k per-generation cap.

OlympiadBench Physics (Qwen3, 4k per-generation cap).

A.10 Per-Family Pathology Decomposition

Decomposing DRG's gain over the greedy baseline B0 as (trigger rate) × (per-trigger retry-fix rate − per-trigger damage rate) on MATH-500 across all five models:

Model	Trig %	W→C	C→W	Fix/trig	DRG−B0
Qwen3-8B	71.6%	62	1	17.3%	+12.2pp
Qwen3-14B	62.6%	51	0	16.3%	+10.2pp
DS-R1-Qwen	50.4%	41	1	16.3%	+8.0pp
DS-R1-Llama	52.6%	35	2	13.3%	+6.6pp
Ministral	31.8%	20	2	12.6%	+3.6pp

Table 13: Per-family pathology decomposition on MATH-500 (canonical diagnostic run; Table 1 reports 5-seed headline means). DS-R1-Qwen/Llama are 7B/8B. *Per-trigger retry-fix rate is stable (12.6–17.3%) across three training families; trigger rate varies* $2\times$. DRG's gain therefore scales mainly with how often the baseline gets stuck. The net product (trigger rate) × (fix rate minus rare C→W damage) reproduces DRG−B0 within 1pp.

A.11 Previous-Attempt Context Window Sweep

We sweep the previous-attempt context window on Qwen3-8B MATH-500 over $\{0, 300, 600, 1200, 2400\}$ characters with 3 retry seeds at each window (B0 deterministic; SC-2 5-seed mean = 64.7%). All other DRG hyperparameters (trigger thresholds, gate, sampling) are held at the defaults from Section 3.

chars	DRG	tok	no-ans	W→C	C→W
0	64.1 ± 0.6	5,575	33.5%	23.0	0.7
300	67.1 ± 0.7	5,326	30.5%	40.0	2.3
600	68.9 ± 0.3	5,231	28.4%	48.7	2.0
1200*	$\mathbf{71.8 \pm 1.1}$	**5,016**	**25.5%**	**63.0**	**2.0**
2400	75.1 ± 0.5	4,525	21.9%	78.7	1.0

Table 14: Context-window sensitivity on Qwen3-8B MATH-500 (3 retry seeds per window). All nonzero context windows outperform the zero-context endpoint, no-answer rate falls, and C→W flips remain rare (≤2.3). The 1200-character row is the fixed default used for all headline results.

Interpretation. This is a sensitivity check, not a tuning protocol: all non-zero context windows outperform the zero-context endpoint, confirming the role of previous-attempt context. We retain 1200 as the default because it was fixed before headline benchmark evaluation. The 2400-character row improves accuracy and generated-token cost on Qwen3-8B, so we treat the headline as a conservative pre-specified operating point and leave context-window selection to held-out calibration.

A.12 Length-Calibration Robustness

The length signal is the only distributional parameter, so we test whether the transductive percentile estimate (computed over the evaluated stream) leaks information relative to a deployment-realistic calibration. We re-derive the 85th-percentile length threshold under two causal alternatives — a prefix-online estimate (growing window, 50-problem warmup) and a 5-fold held-out split — and recompute every trigger decision on the two frontier models (MATH-500, canonical diagnostic run).

Model	Thr.	Cap-hit	Online flips	5-fold flips
Qwen3-8B	4096	39.0%	0 / 450	0 / 500
DS-R1-Qwen-7B	4096	25.0%	0 / 450	0 / 500

Table 15: Length-calibration robustness on the two frontier models. "Thr." is the 85th-percentile token threshold; "Cap-hit" is the fraction of B0 traces at the 4096 cap; "flips" counts trigger decisions that change versus the transductive estimate.

On both models the 85th percentile pins to the 4096-token cap, because the cap-hit rate exceeds the complementary 15% tail. As a result, prefix-online and 5-fold held-out calibration produce the *identical* trigger set — $0/450$ post-warmup online decisions and $0/500$ held-out decisions change —

so the transductive estimate carries no measurable advantage in the budget-pressure regime where DRG operates. (Reconstructing the threshold rule reproduces the stored trigger decisions $500/500$.) Only 4 (Qwen3-8B) and 6 (DS-R1-Qwen-7B) problems trigger on length alone without a repetition or stall backup, bounding any residual sensitivity. Below 15% cap-hit the threshold would fall under the cap and calibration choice could matter; we therefore scope this null result to the high-pressure cells.

A.13 Latency vs. Throughput

Our compute accounting reports *generated tokens*, which is the binding cost under throughput-bound (batched) serving. Under latency-optimized serving the relevant quantity is instead the critical path: SC-K issues its K samples in parallel (latency $=$ the longest sample), whereas DRG runs baseline then retry sequentially (latency $=$ baseline $+$ retry). Table 16 contrasts the two on the frontier models (canonical diagnostic run, mean tokens/problem, batch-1 parallel-replica assumption).

Model	Tok./SC2	Lat./SC2	Lat./SC8
Qwen3-8B	0.81×	1.56×	1.45×
DS-R1-Qwen-7B	0.70×	1.26×	1.08×

Table 16: Generated-token ratio (Tok.; throughput proxy) vs. critical-path latency ratio (Lat.; parallel SC-K serving). DRG's token advantage does not transfer to latency because its retry is sequential.

DRG's 0.70–$0.81\times$ generated-token advantage becomes a 1.26–$1.56\times$ latency *disadvantage* once SC-K samples are run on parallel replicas, because the retry cannot start until detection on the baseline completes. The token-cost win is therefore a throughput/serving-cost win, not a wall-clock-latency win; under batched serving (where samples contend for the same hardware) generated tokens again track latency and the DRG advantage is restored. We report the metric (generated tokens) appropriate to throughput-bound deployment and disclose the latency regime here.

A.14 What DRG Fixes: Finalization vs. Reasoning

To separate finalization recovery from substantive reasoning correction, we classify every W→C fix by whether the baseline produced an extractable answer at all.

Model	W→C	Finalization	Reasoning	C→W
Qwen3-8B	62	62 (100%)	0 (0%)	1
DS-R1-Qwen-7B	41	41 (100%)	0 (0%)	1

Table 17: Fix-type breakdown (canonical diagnostic run). "Finalization" = baseline had no extractable answer; "Reasoning" = baseline produced a wrong extractable answer that the retry corrected.

All observed W→C fixes in these two runs occur on no-extractable-answer baselines, confirming that DRG targets budget-exhausted finalization failures rather than wrong-answer correction (and no-answer-only triggering reproduces full DRG accuracy on all five models, Table 5). The matched fresh-retry baseline shows this is not merely formatting: adding previous-attempt context recovers $+6.8$/$+6.9$pp beyond a no-context restart (Table 2), i.e. the useful operation is not just asking for a final answer but restarting with information from the failed attempt. This is consistent with the pathology gate's ≥ 2-signal requirement, which finalized-but-wrong traces rarely satisfy.

A.15 Forced-Finalization Control

Does merely *asking* the model to finalize recover the no-answer failures, without a full restart? We test a budget-forcing / forced-termination control: on a no-answer baseline we inject an end-of-thinking marker plus an answer cue onto the final 800 characters of A_0 and generate a short ≤ 256-token continuation, accepting it only when it yields an extractable answer. This is the fair no-answer-only reading (single seed) — it acts only on already-failed baselines and so cannot flip a correct A_0 (unlike an accept-all-disagreements gate, which causes avoidable C→W on DS-R1).

Model	B0 acc.	F-fin	DRG	Fix/NA	Resid
Qwen3-8B	59.6	64.6	**72.2**	25/191	32.6%
Qwen3-14B	64.6	66.0	**75.0**	7/168	32.0%
DS-R1-Qwen-7B	71.8	71.8	**79.4**	0/114	22.8%

Table 18: **Forced-finalization control** (single seed, MATH-500, no-answer-only). **F-fin**: forced-finalization accuracy; **DRG**: same-seed no-answer DRG; **Fix/NA**: no-answer cases the control corrects out of all no-answer baselines (DRG corrects 63/52/38 of the same pools); **Resid**: residual no-answer rate after the control. Forced finalization at $3{,}155/2{,}990/2{,}200$ tokens/problem.

Forced finalization fixes only 25/7/0 no-answer cases versus 63/52/38 for restart-with-context DRG, and nothing on DS-R1-Qwen-7B. Contin-

uing the stuck prefix — even solely to emit a boxed answer — rarely produces a *correct* one; the useful operation is a fresh restart conditioned on previous-attempt context, consistent with the perturbation ladder (Appendix A.25). A forced-finalize-before-DRG ladder (try the cheap finalization first, fall back to a full restart otherwise) may slightly improve one cell at comparable tokens but is a single-seed diagnostic, not the main mechanism.

A.16 Compute-Matched SC-K Frontier

We derive realized SC-K Pareto frontiers (K=1, 2, 4, 8) by prefix-aggregating SC-8 sample pools — *three independent pools* (seeds 1–3) — on Qwen3-8B, Qwen3-14B, and DS-R1-Qwen-7B (MATH-500). K=2 uses the same first-sample tie-break as the SC-2 baseline; $K \in \{4, 8\}$ use symbolic-equivalence cluster voting. Reporting mean±std over the three pools tests whether the low-token frontier point is an artifact of a single lucky pool; it is not (SC-8 std $\leq$ 0.8pp).

A.17 Cross-Model Consistency (MATH-500, DS-R1 / Mistral)

A.18 W-to-W Failure and Per-Difficulty Breakdown

Full W→W breakdown:

Qualitative patterns in no-answer failures. Examining 30 sampled W→W cases, these are predominantly Level 4–5 problems in intermediate algebra, geometry, and precalculus. The model begins reasoning but exhausts the 4096-token budget before reaching a boxed answer. Many show the model starting a correct approach but running out of space for final computation. Improved answer extraction would likely benefit all methods equally and would not change the relative comparison.

Per-difficulty breakdown (MATH Levels 1–5, Qwen-7B, best trigger combo rep+length). Trigger rate increases monotonically with difficulty in this Qwen-7B MATH-500 breakdown (16% → 78%). Largest accuracy gains are at Levels 2–3 (sufficient capability + budget pressure); Level 1 is essentially unchanged (zero accuracy change with only 7 triggers); Level 5 shows the capability boundary (highest trigger rate but +9.0pp gain, lower than Levels 2–4 — the model triggers frequently but often cannot solve even with a fresh attempt).

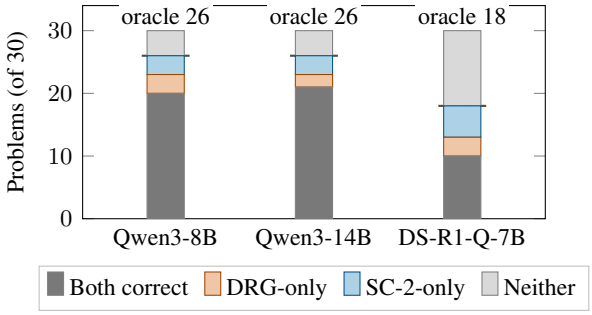

Figure 6: **Oracle complementarity on AIME-32k (single-seed diagnostic, n = 30).** Bars partition 30 problems by DRG/SC-2 outcome on three models; dashed ticks mark the any-path oracle. **Takeaway (hypothesis-generating):** the +6.7–10pp headroom appears to lie in the disagreement set, suggesting future routers must decide not only when to retry, but which successful path to trust.

A.19 High-Budget (32k) Cross-Benchmark Diagnostics

The 32k AIME/AMC cells are *single-seed* probes (n=30/40) used to illustrate operating regimes, not for fine-grained claims. Table 23 lists them with the exploratory regime taxonomy.

At 32k, these single-seed probes illustrate possible boundary behavior as pressure falls: several AMC cells converge near saturation or favor SC-2 by a few problems. Single-seed AIME-32k diagnostics suggest DRG and SC-2 reach comparable accuracy on partially disjoint problem sets, leaving oracle headroom in the disagreement set (Figure 6) that neither DRG nor Deploy exploits — a routing question, not a no-answer one. Because these cells are single-seed at n=30/40 (±3pp ≈ one problem), we read them as regime illustrations rather than claims.

A.20 DRG–SC-2 Overlap Counts

Full overlap counts referenced in Section 4.4 and Figure 6:

A.21 Deploy Pipeline Routing Outcomes (MATH-500)

Per-model Deploy outcomes referenced in Section 4.2:

A.22 Trigger-Selection Controls on Qwen3 Family (Single Seed)

Single-seed controls on Qwen3-8B and Qwen3-14B referenced in Section 4.4:

	Qwen3-8B		Qwen3-14B		DS-R1-Qwen-7B	
Method	**Acc**	**Tok**	**Acc**	**Tok**	**Acc**	**Tok**
SC-1 (sampled)	59.3±0.7	3,076	64.3±0.1	2,895	73.3±0.7	2,055
SC-2	64.2±0.6	6,148	68.6±0.7	5,780	77.9±0.6	4,126
SC-4	68.0±0.3	12,330	71.7±0.5	11,591	82.5±0.5	8,220
SC-8	71.7±0.4	24,675	74.9±0.4	23,214	85.7±0.8	16,391
DRG	**71.6**	**5,001**	**75.3**	**4,615**	**79.1**	**3,340**

Table 19: **Multi-pool SC-K frontier on MATH-500** (3 independent SC-8 pools, seeds 1–3; SC-K accuracy is mean±std over pools, DRG from Table 1). SC-8 variance is small (std ≤ 0.8pp), so the frontier is not a single-pool artifact. On *both* Qwen3 models DRG matches or exceeds the SC-8 *mean* at $0.20\times$ SC-8 tokens and dominates SC-2/SC-4 in both axes; on DS-R1-Qwen-7B DRG dominates the SC-2 point, while SC-4/8 buy higher accuracy at $\sim 2.5\times/\sim 5\times$ DRG tokens.

Property	**Qwen-7B**	**Llama-8B**	**Ministral**
Trigger rate	50.4%	52.6%	31.8%
Accuracy gain over B0	+7.3pp	+7.1pp	+3.0pp
Δ over SC-2	-0.3pp	$+3.3$pp	-2.4pp
Token ratio vs SC-2	$0.69\times$	$0.72\times$	$0.71\times$
C→W count (mean)	1.8	2.6	1.6
Control 2 Cohen's d	13.98	6.81	5.26

Table 20: Cross-model consistency of DRG behavior across the three non-Qwen3 MATH-500 models. Gains over B0 are 5-seed means (Table 1); the canonical single-run decomposition in Table 13 differs by $\leq$ 1pp. Trigger rate and improvement over B0 decrease as stuck-failure headroom shrinks; standalone accuracy relative to SC-2 depends on regime, while correctness preservation remains high.

Failure Reason	**Count**	**%**
Triggered, disagree, accept no-answer	61	61.0%
Not triggered, keep baseline wrong	17	17.0%
Triggered, disagree, low-pathology keep wrong	9	9.0%
Triggered, agree, keep wrong	8	8.0%
Triggered, disagree, accept wrong	5	5.0%

Table 21: W→W failure breakdown (Qwen-7B MATH-500). The dominant mode is the gate correctly accepting a retry that fails to produce an extractable answer within budget — a formatting ceiling shared across methods.

Level	**N**	**B0 acc.**	**DRG**	**Δ**	**Trig.**	**Rate**
1 (AMC 8)	43	93.0%	93.0%	+0.0	7	16.3%
2	90	84.4%	92.2%	+7.8	24	26.7%
3	105	77.1%	89.5%	+12.4	45	42.9%
4	128	69.5%	79.7%	+10.2	72	56.3%
5 (AIME)	134	54.5%	63.4%	+9.0	104	77.6%

Table 22: Per-difficulty performance on Qwen-7B MATH-500.

Bench.	**Model**	**B0 acc.**	**DRG**	**SC-2**	**Oracle**	**Reg.**
AIME	Qwen3-8B	66.7	76.7	76.7	**86.7**	X
AIME	Qwen3-14B	73.3	76.7	80.0	**86.7**	X
AIME	Qwen-7B	36.7	43.3	50.0	**60.0**	X
AIME	Ministral-14B	40.0	**50.0**	43.3	53.3	R
AMC	Qwen3-8B	92.5	95.0	100.0	100.0	C
AMC	Qwen3-14B	95.0	97.5	100.0	100.0	C
AMC	Ministral-14B	80.0	**85.0**	82.5	87.5	V
AMC	Qwen-7B	75.0	80.0	87.5	90.0	V

Table 23: High-budget 32k AIME/AMC accuracy (%), single-seed diagnostic probes. Regime labels are exploratory: R=recovery, C=convergence, V=convergence leaning SC-2, X=complementarity.

Setting	**Both**	**DRG**	**SC2**	**None**	**Hd.**
M4 Q3-8	307	52	12	129	+2.4
M4 Q7	369	30	28	73	+5.6
A4 Q3-8	1	4	0	25	0.0
A4 Q3-14	1	3	1	25	+3.3
A4 Min	11	3	2	14	+6.7
C4 Q3-8	14	5	0	21	0.0
C4 Q3-14	14	6	0	20	0.0
C4 Min	31	2	0	7	0.0
A32 Q3-8	20	3	3	4	**+10.0**
A32 Q3-14	21	2	3	4	**+6.7**
A32 Q7	10	3	5	12	**+10.0**
A32 Min	12	3	1	14	+3.3
C32 Q3-8	38	0	2	0	0.0
C32 Q3-14	39	0	1	0	0.0
C32 Q7	31	1	4	4	+2.5
C32 Min	32	2	1	5	+2.5
OlyP Q3-8	8	5	2	221	+0.9
OlyP Q3-14	10	7	2	217	+0.9

Table 24: Oracle complementarity between DRG and SC-2. Hd. is headroom in pp over the best standalone path. M=MATH-500, A=AIME, C=AMC, OlyP=OlympiadBench Physics; Q3=Qwen3, Q7=DS-R1-Qwen-7B, Min=Ministral-14B.

Model	Deploy	Oracle	Rescue	Worse
Qwen3-8B	73.8±0.5%	74.0±0.5%	11.0	0
Qwen3-14B	76.8±0.4%	76.8±0.4%	7.8	0
Qwen-7B	82.5±0.4%	84.7±0.4%	17.0	0
Llama-8B	79.6±0.5%	84.5±0.3%	9.6	0
Min-14B	90.0±0.5%	91.6±0.2%	11.0	0

Table 25: Deploy pipeline routing outcomes on MATH-500 (4k per-generation cap), reported as 5-seed means over retry seeds 1–5. Rescues = average SC-2 fallbacks that fix DRG no-answer failures; worsens are zero on all models.

Model	Selective	Random	Universal
Qwen3-8B	71.8%	68.8% (−3.0)	71.8%
Qwen3-14B	74.8%	72.0% (−2.8)	74.8%

Table 26: Random rate is matched to the real selective rate. Universal triggering matches selective in accuracy but uses more retry compute; random underperforms by 2.8–3.0pp, mirroring the 50-seed pattern from the DS-R1 / Mistral models.

A.23 Full Control Details

A.23.1 Control 4: Correctness Preservation

Model	$n_{\text{triggered}}$	C→W	Preservation
Qwen3-8B	358	1	99.4%
Qwen3-14B	313	0	100.0%
Qwen-7B	252	1	>97%
Llama-8B	263	4	97.2%
Ministral-14B	161	2	97.9%

Table 27: Correctness preservation under DRG gating (canonical run). C→W counts originally-correct A_0 outputs flipped to wrong by the gate; preservation is over originally-correct triggered traces. All five models are $\geq 97\%$.

A.23.2 Control 2: Random Trigger Distribution (50 Seeds)

Model	Real	Rand. $\mu \pm \sigma$	95th	d	p
Qwen-7B	78.2%	72.3% ± 0.42%	72.9%	13.98	0.020
Llama-8B	78.0%	74.5% ± 0.51%	75.4%	6.81	0.020
Min-14B	89.0%	86.2% ± 0.52%	87.0%	5.26	0.020

Table 28: Control 2 random trigger distribution (50 seeds).

Operator	Keep A0	Low path	Accept A1	Fix/Acc.
NR	133	11	108	38.9%
Try	130	13	109	31.2%
Step	131	12	109	32.1%
Neg.	123	19	110	27.3%
A1	133	12	107	38.3%
B1	130	14	108	38.0%
WSC-style	132	12	108	33.3%
Paraphrase	129	16	107	34.6%

Table 29: Gate outcome breakdown by operator.

Operator	Both	NR-only	OP-only	Neither	p
Try	156	13	4	79	0.052
Step	157	12	4	79	0.080
Neg.	148	21	7	76	0.014
A1	158	11	11	72	0.831
B1	157	12	11	72	1.000
WSC-style	157	12	6	77	0.239
Paraphrase	153	16	11	72	0.441

Table 30: McNemar contingency summary versus NR.

A.23.3 Gate Breakdown by Operator (Qwen-7B, Triggered Subset $n = 252$)

A.23.4 McNemar Contingency Tables (vs NR, Triggered Subset)

Note: NR-only column denotes problems where NR fixes but the operator does not. OP-only denotes problems where operator fixes but NR does not. Negative feedback has the most asymmetric table (21 vs 7), explaining its significant McNemar result.

A.23.5 Fix Overlap Statistics

- Problems fixed by all 8 operators: **9** (universally rescuable)
- Problems fixed by exactly 1 operator: **18** (operator-specific)
- Union of all operator fixes: **70** (oracle upper bound)
- NR fixes: **42** (captures 60% of oracle pool)

A.23.6 Jaccard Similarity Matrix (Fix Sets)

A.23.7 Token Breakdown by Operator

Note: token delta is negative because retries on stuck problems produce shorter outputs than the original stuck trace. All operators save tokens globally.

A.24 Oracle Gating Details

Full per-model oracle table referenced in Section 4.3:

	NR	Try	Step	Neg	A1	B1	WSC	Para
NR	1.00	0.65	0.67	0.47	0.60	0.57	0.63	0.49
Try	0.65	1.00	0.57	0.42	0.53	0.50	0.67	0.45
Step	0.67	0.57	1.00	0.44	0.52	0.49	0.58	0.50
Neg	0.47	0.42	0.44	1.00	0.48	0.45	0.54	0.46
A1	0.60	0.53	0.52	0.48	1.00	0.55	0.64	0.47
B1	0.57	0.50	0.49	0.45	0.55	1.00	0.57	0.53
WSC	0.63	0.67	0.58	0.54	0.64	0.57	1.00	0.55
Para	0.49	0.45	0.50	0.46	0.47	0.53	0.55	1.00

Table 31: Jaccard similarity across operator fix sets.

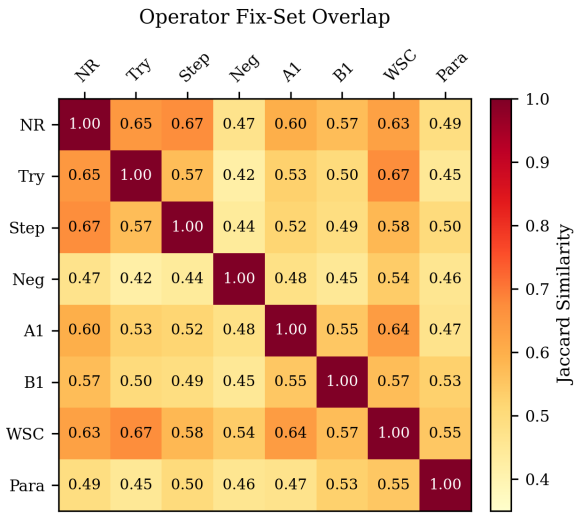

Figure 7: Jaccard similarity of fix sets across 8 retry operators (Qwen-7B, triggered subset). Negative feedback (Neg) shows the lowest overlap with all other operators, consistent with its significantly lower fix rate. NR, A1, and Step show the highest mutual overlap.

Operator	Δ tok.	Acc. retry	Rej. retry	Global Δ
NR	−919	2,269	3,359	−463
Try	−844	2,343	3,448	−425
Step	−844	2,344	3,772	−426
Neg.	−780	2,389	3,373	−393
A1	−892	2,298	3,188	−450
B1	−927	2,261	3,275	−467
WSC-style	−847	2,330	3,203	−427
Paraphrase	−850	2,327	3,313	−429

Table 32: Token breakdown by operator.

Model	B0 acc.	Gate	Oracle	Head	T1	T3
Qwen-7B	71.8%	80.0%	80.8%	+0.8	4	0
Llama-8B	70.6%	78.0%	79.0%	+1.0	1	4
Ministral-14B	84.8%	89.0%	89.4%	+0.4	0	2

Table 33: Gate performance vs. oracle. T1 = missed-fix; T3 = C→W acceptance.

Level	Acc	Description
Baseline	71.8%	Single greedy generation
DRG (NR + gate)	80.0%	Our method (1 retry)
8-op oracle (gated)	85.8%	Perfect operator, current gate
8-op oracle (retry)	87.2%	Perfect operator + perfect gate

Table 34: Oracle compute-accuracy ladder on Qwen-7B.

8-operator oracle on Qwen-7B.

A.24.1 Triggered-Subset Oracle Breakdown

Model	n_t	B0 acc.	Gate	Oracle	Head	T1	T3
Q7B	252	50.8%	67.1%	68.7%	+1.6pp	4	0
Q1.5B	413	8.5%	15.0%	15.7%	+0.7pp	2	1
L8B	263	53.6%	67.7%	69.6%	+1.9pp	1	4
Min	161	60.2%	73.3%	74.5%	+1.2pp	0	2

Table 35: Triggered-subset oracle breakdown. Q7B=DS-R1-Qwen-7B, Q1.5B=DS-R1-Qwen-1.5B, L8B=DS-R1-Llama-8B, Min=Ministral-14B.

Type 1: baseline wrong, retry correct, gate kept baseline (missed fix). Type 3: baseline correct, retry wrong, gate accepted retry (C→W).

A.24.2 8-Operator Oracle (Qwen-7B)

Metric	Global	Triggered
Baseline	71.8%	50.8%
Best single gate (NR)	80.0%	67.1%
8-op oracle (gated finals)	85.8%	78.6%
8-op oracle (any retry correct)	87.2%	81.3%
Headroom vs best gate (gated oracle)	+5.8pp	+11.5pp
Headroom vs best gate (retry oracle)	+7.2pp	+14.3pp

Table 36: 8-operator oracle summary on Qwen-7B.

Gate miss breakdown (best single gate vs 8-op retry-oracle, triggered): 36 misses out of 205 oracle-correct problems (17.6%). Of these, 35 are Type 1 (missed fix) and 1 is Type 3 (C→W).

A.25 Perturbation Ladder Details

Full perturbation ladder referenced in Section 4.3:

Perturb.	Ctx	Acc	Trig.	Fix	C→W	Budget
B0	100%	71.8%	50.8%	–	–	4,096
Trunc $\frac{1}{3}$	67%	73.8%	54.8%	13	3	1,935
Trunc $\frac{1}{2}$	50%	74.6%	56.3%	15	1	2,483
Problem-only	0%	75.8%	58.7%	23	3	4,096
Restart	Fresh	80.0%	67.1%	42	1	4,096

Table 37: Perturbation ladder, Qwen-7B MATH-500 triggered subset.

Confound analysis: truncate-to-problem vs. full restart. On the 135 problems where both methods produce extractable answers: retry correctness 92.6% (truncate) vs. 94.8% (restart); final accuracy after gating 91.9% vs. 93.3%; head-to-head wins 2 (restart) / 0 (truncate) / 133 ties.

A.25.1 Budget-Normalized View

Perturb.	Acc	Mean tok.	Acc/1k tok.
Baseline	71.8%	2,141	0.34
Trunc last third	73.8%	2,228	0.33
Trunc to half	74.6%	2,185	0.34
Trunc to problem	75.8%	2,107	0.36
Restart	80.0%	1,678	0.48

Table 38: Budget-normalized perturbation view.

A.25.2 Diagnostic: No-Answer and Budget Exhaustion Rates

Perturb.	No-ans.	Budget ex.	Rate
Trunc $\frac{1}{3}$	117	138	46% / 55%
Trunc $\frac{1}{2}$	106	122	42% / 48%
Problem-only	106	116	42% / 46%
Restart	70	76	28% / 30%

Table 39: No-answer and budget exhaustion diagnostics.

The high no-answer rates for truncation methods reflect the disruption of the model's output structure when continuing from a truncated prefix. Full restart, generating from a clean prompt template, produces extractable answers more reliably.

A.25.3 No-Answer Overlap (Truncate-to-problem vs Full Restart)

Category	Count
P3 no-answer	106
P4 no-answer	70
Both no-answer	59
P3 no-answer, P4 has answer	47
P3 has answer, P4 no-answer	11

Table 40: No-answer overlap between truncate-to-problem and full restart.

A.26 Operator Strings and Descriptions

NR (no prompt): No additional text. Fresh generation from original problem.

Try again: "Try again."

Step-by-step: "Solve this problem step by step."

Negative feedback: "Your previous answer was wrong. Start over and solve this problem."

A1 regrounding: Structured regrounding prompt: restate the problem, identify givens and unknowns, then solve step by step.

B1 strategy switch: Structured strategy-switch prompt: try a different representation or solution strategy before deriving the final answer.

WSC-style: "I realize I can find a clearer solution. Let me approach this problem with a fresh perspective and careful reasoning."

Paraphrase: Rule-based transformation of problem text: clause reordering, synonym swaps from fixed dictionary, sentence reordering. Applied to retry input only (not baseline). All LaTeX preserved. Ground truth verified unchanged.

A.27 Cochran's Q Test Details

Cochran's Q uses 252 triggered problems as subjects and 8 operator treatments ($df = 7$), yielding $Q = 14.517$ and $p = 0.043$. The significant Q rejects strict equivalence across all 8 operators. Post-hoc pairwise McNemar tests (Table 30) localize the difference to negative feedback ($p = 0.014$ vs NR). After excluding negative feedback, the remaining 7 operators show no significant pairwise differences.

A.28 Pressure Diagnostics and Extraction Robustness

Cross-family pressure regression (primary; strict extraction). Regressing the DRG−SC-2 gain (Table 1 values) on the B0 no-answer rate across the five MATH-500 4k cells (Table 41) gives slope $+0.353$pp/pp, intercept -5.78, $R^2 = 0.84$, Spearman $\rho = 0.80$, and a zero-crossing at 16.4% no-answer. This descriptive fit supports no-answer pressure as the main operating signal, but not the only factor: the two DeepSeek models sit at similar pressure ($\sim$21–23%) yet differ by 3.6pp, so family, sampling variance, and model-specific behavior still matter.

Model	Family	B0 no-ans %	Gain
Qwen3-8B	Qwen3	38.2	+6.9
Qwen3-14B	Qwen3	33.6	+7.7
DS-R1-Qwen-7B	DeepSeek	22.8	−0.3
DS-R1-Llama-8B	DeepSeek	21.0	+3.3
Ministral-14B	Mistral	9.2	−2.4

Table 41: Cross-family pressure diagnostic inputs (strict, 4k). Gain = DRG−SC-2 (pp). Descriptive fit: slope $+0.35$, $R^2 = 0.84$, $\rho = 0.80$, zero-crossing 16.4%.

Uniform permissive-extraction audit (Qwen3-8B). We re-extract every method's saved generations with one permissive extractor (last `\boxed{}`,

then "final answer is X"/"answer: X", then trailing equation result). SC-K was re-run with `-save_sample_texts` (single pool); DRG is 3-seed. Under uniform extraction (Table 42) DRG's edge over SC-2 persists ($+5.7$pp) and SC-8 reaches comparable accuracy at much higher token cost; absolute accuracies of all methods rise because the strict boxed-answer evaluator misses unboxed answers. Forced finalization gains nothing under permissive extraction ($+0.0$/$+0.4$/$+0.0$pp on the three models), confirming it is a reformatting artifact, not recovery.

Method (Qwen3-8B)	Strict	Permissive
B0	59.6	73.6
SC-2	64.4	75.0
SC-8	72.0	80.2
DRG	**71.8**	**80.7**

Table 42: Uniform permissive-extraction audit, Qwen3-8B MATH-500 (%). DRG$-$SC-2 holds (strict $+7.4$, permissive $+5.7$); SC-8 reaches comparable accuracy at $5\times$ DRG tokens. SC-K single-pool with saved texts; DRG 3-seed.

Within-Qwen3 cap-sweep (secondary; both extractors). Holding the model fixed and varying the cap (Table 43) varies pressure causally. Under permissive extraction, listed by cap $2k/4k/8k$, the DRG$-$SC-2 gains are $+5.6$/$+5.0$/$+2.6$; equivalently, the gain rises as pressure increases, i.e., as the cap tightens from $8k$ to $2k$. Under strict extraction the gain dips at $2k$ ($+3.6$) because high-pressure traces rarely emit a boxed final answer, so the strict evaluator undercounts the very recoveries DRG produces — the finalization failure surfacing in the extractor. The strict gains remain positive but are not monotonic, illustrating why the permissive audit is a robustness check rather than the primary metric.

Cap	B0 no-ans %	Strict gain	Perm. gain
2048	77.2	$+3.6$	$+5.6$
4096	38.2	$+7.2$	$+5.0$
8192	18.6	$+3.4$	$+2.6$

Table 43: Qwen3-8B cap-sweep, DRG$-$SC-2 gain (pp). Permissive is monotonic in pressure; the strict $2k$ dip reflects boxing failure under extreme pressure (the finalization artifact itself).