CHANGING THE CANDIDATES, NOT THE VOTE: A REPAIR PORTFOLIO FOR SELF-CONSISTENCY UNDER A PER-CALL TOKEN CAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-consistency (SC) votes over sampled reasoning chains, but under a hard per-call token cap many chains never finish, and a system that discards unfinished generations has nothing to vote for. We study an alternative at an equal number of calls: replace three of eight SC samples by a greedy chain that does not vote and two repairs of it, and return the SC plurality when any SC sample answers and the repair plurality otherwise. With Qwen3.5 models on MATH Level 4–5, this frozen configuration raises accuracy over SC@8 by 10.0 percentage points with 95% confidence interval $[5.8, 14.5]$ in an exploratory study at a 4,096-token cap, by 3.7 $[2.1, 5.6]$ in a pre-registered evaluation of a 475-question archive, and by 9.4 $[5.5, 13.8]$ and 11.7 $[6.3, 17.7]$ in a pre-registered replication on 100 unused questions at caps of 4,096 and 16,384 tokens. The gain comes from answer supply: it arises where no SC chain finishes, because the repairs finish more often. It persists when answers are recovered from unfinished chains and when SC is sampled at the repairs' temperature or with a budget instruction in every prompt, and the 15-call portfolio exceeds SC@16 at the same cap in the original study and at both prospective caps.

1 INTRODUCTION

Self-consistency samples several chain-of-thought traces (Wei et al., 2022; Wang et al., 2023) and returns the most frequent final answer. Two conditions must hold for this to work: the correct answer must be the model's modal answer, and the samples must return answers at all. The first failure, plurality inversion, is understood (Chen et al., 2024). The second is the subject of this paper. Long-reasoning models often re-verify an answer they have already reached (Chen et al., 2025); under a fixed per-call budget such chains are stopped before they terminate, often after they have written a candidate answer, and a system that counts only finished generations abstains. At a 4,096-token budget, Qwen3.5-27B leaves all eight SC chains unfinished on a third of the hard MATH records of our original study. Adding samples from the same process closes this abstention floor slowly, as Section 3 shows: pooling 24 SC chains at the same cap still leaves a quarter of these questions without any answer. Changing the candidates is a natural alternative, and the gains we report come from answer supply, not from correcting inverted pluralities.

The cap is an external constraint of the deployment, for example a serving limit on each call, and so is the model's reasoning mode, which we assume to be its thinking mode; we keep four quantities apart throughout: correctness of a stated answer, termination within the budget, extraction of the stated answer, and abstention by the return policy. Our contributions are:

- a measurement of this termination failure in Sections 3 and 6.4: SC abstains on a large share of hard questions, and pooled SC samples close the floor as slowly as a simple model predicts;
- an *empirical portfolio result under a hard per-call cap and a specified return policy*: at an equal number of calls, replacing three SC samples by a non-voting greedy chain and two repairs of it improves accuracy, confirmed in the two pre-registered evaluations of Sections 6.2 and 6.3;
- a mechanism with registered controls in Section 6.4: the repairs finish more often and supply answers where SC supplies none; matching their temperature and final-line instruction, or adding a

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

budget instruction to every prompt, leaves the advantage in place, and it can be estimated before adoption from the SC no-answer rate.

The claim concerns the candidates: on the audited records of the three main evaluations the rule reduces to "use SC if it answers, else the repairs", as Section 4 shows, so aggregation rules, entropy-based step selection and the detection of inverted questions lie outside its scope.

The evidence has three parts of different standing, which we report separately. First, the *original study* of 100 questions under three seeds is exploratory: the eight-call configuration was specified after its per-arm results were known, so its intervals are conditional on that choice. Second, an *archival evaluation* on a separately collected 500-question portfolio was pre-registered, with the configuration, primary population, endpoints and decision rule fixed before any current-evaluator output existed; the archive's questions were used in earlier, legacy-evaluated analyses, which we disclose with every number. Third, a *prospective replication* was registered before any of its generations existed: 100 questions never used in this project, a second model size, two budgets.

2 Related Work

Voting and repeated calls. The jury theorem (Condorcet, 1785) underlies SC; Chen et al. (2024) characterise query difficulty by the gap between the most likely wrong and the correct answer and show that voting accuracy can first rise and then fall with the number of calls. Feng et al. (2026) study SC as mode estimation with adaptive sampling, and Cordero-Encinar & Duncan (2025) certify that voting has recovered the mode; both target the mode, which is wrong under inversion and undefined when no sample finishes. Wang et al. (2024) use agreement statistics as confidence signals.

Changing the candidates. Mirror-Consistency (Huang et al., 2024b) uses disagreement among generations to prompt reflection and regeneration; prompt-diversity methods (Mu et al., 2024; Hu et al., 2025), universal self-consistency and progressive hints (Chen et al., 2023; Zheng et al., 2023), and self-refinement (Madaan et al., 2023; Huang et al., 2024a) change how candidates are produced or selected. DiVeRSe (Li et al., 2023) combines diverse prompts with a trained step-aware verifier that weights candidates, and X-of-Thoughts (Liu et al., 2023) switches to another reasoning method when an external check fails; both assume that candidates finish, whereas our repairs address pools that contain no answer. Test-time compute and repeated-sampling studies (Snell et al., 2025; Brown et al., 2024) and verifiers (Lightman et al., 2024) ask how to allocate inference-time computation and select candidates; we use no trained verifier, and verifier-based selection is complementary to the change of candidates studied here. The closest mechanisms revise candidates before voting. PACER (Zhang et al., 2026) builds a consensus packet from the answers of a candidate pool and lets each trace revise itself once against it; Mirror-Consistency reflects on disagreement among finished generations; Snell et al. (2025) trade sequential revisions by a fine-tuned revision model against parallel samples. All three draw their revision signal from answers already present in the pool. Our setting is the complementary one in which the pool contains no answer: the repairs are conditioned on one greedy attempt, not on a consensus, use the base model without training, and, on the audited records of the three main evaluations, matter only when no SC chain finishes, so that there they can be deferred; a comparison with consensus-conditioned revision at our call count is outside this evaluation.

Reasoning budgets. Budget forcing (Muennighoff et al., 2025), budget-aware prompting (Han et al., 2025) and skipping explicit thinking (Ma et al., 2025) change how a single chain uses its budget. We take the cap and the decoding of each call as given and ask what can be done when many chains exhaust the cap; Section 6.4 shows that a budget instruction raises SC's answer supply and that the repairs add to it, and Section 6.5 evaluates recovering answers stated in unfinished chains.

3 Background: When Voting Cannot Recover the Correct Answer

This section records two limits of answer voting that motivate changing the candidates; proofs are in Appendix C. For a question q, each SC sample returns an answer X_i drawn from a categorical distribution over the reference answer y and wrong answers $a_1, \ldots, a_{C-1}$, with probabilities p_q and $p_{q1}, \ldots, p_{q,C-1}$; as an explicit modelling assumption throughout, samples are conditionally independent and identically distributed given q. SC returns the plurality answer among $X_1, \ldots, X_k$.

Definition 1: Plurality gap and inversion**.** *Let $p_q^* = \max_j p_{qj}$ be the probability of the most likely wrong answer. The plurality gap is $\Delta_q = p_q - p_q^*$, and q is* plurality-inverted *when $\Delta_q < 0$.*

Plurality inversion is stricter than the binary threshold $p_q < 0.5$: every inverted question has $p_q < p_q^* \leq 1 - p_q$, hence $p_q < 0.5$, but $p_q = 0.4$ with wrong answers $(0.2, 0.2, 0.2)$ has $\Delta_q = 0.2 > 0$. Chen et al. (2024) characterise query difficulty through the same probability difference with the opposite sign and show that voting accuracy can first rise and then fall with the number of calls; Theorem 1 states this phenomenon for plurality voting in terms of the inversion rate and adds a pairwise exponential bound.

Theorem 1: Non-monotone plurality SC scaling**.** *Let D be a distribution over questions, or equivalently over per-question answer distributions, $r_P = \Pr_D(\Delta_q < 0)$ the plurality inversion rate, and $A(k) = \Pr_{q\sim D}(\text{SC with } k \text{ samples returns } y)$. Assume $\Pr_D(\Delta_q = 0) = 0$. Then:*

1. *$\lim_{k\to\infty} A(k) = 1 - r_P$.*
2. *If $A(k_0) > 1 - r_P$ for some finite k_0, then A is not monotone non-decreasing: $A(k') < A(k_0)$ for some $k' > k_0$, and A attains its global maximum at a finite k^*.*
3. *For an inverted question, let $\rho_q = p_q + p_q^*$, $\pi_q = p_q/\rho_q < \frac{1}{2}$, and $D_q = D_{\mathrm{KL}}(\frac{1}{2} \,\|\, \pi_q)$, with D_{KL} the Kullback–Leibler divergence of Bernoulli laws. The probability that y ties or beats a fixed most likely wrong answer a^* satisfies*

$$\Pr(n_y \geq n_{a^*}) \;\leq\; \left(1 - \rho_q + \rho_q e^{-D_q}\right)^k, \tag{1}$$

so it decays exponentially at rate $-\log(1 - \rho_q + \rho_q e^{-D_q}) \;\geq\; \rho_q(1 - e^{-D_q})$. Because SC can return y only if y ties or beats a^*, the same bound holds for the probability that SC is correct, whatever the tie-breaking rule.

Part 1 says that SC converges to the model's modal answer, not to the correct one; eventual incorrectness on an inverted question does not require the wrong mode to be unique, although a fixed wrong winner does. Part 2 turns a single measurable quantity into a sufficient condition for a finite maximiser, possibly at $k = 1$: whenever some finite budget already beats the modal-answer limit, adding samples must eventually hurt. Part 3 makes the decay per question explicit. For example, for $p_q = 0.05$ with wrong answers $(0.45, 0.25, 0.25)$ the exponent in equation 1 is 0.223 per sample.

Answer-only rules. The bound is not an artifact of plurality counting: it extends to every aggregator that returns one of the answers with the largest empirical count, whatever its tie rule, weighting or confidence threshold, as Proposition 2 in Appendix C shows; candidates with a different answer distribution are one way out, and information beyond the answers is another.

Unfinished chains. The second failure mode, samples that return no answer, fits the same framework. Suppose each sample is independently unfinished with probability $\tau_q \in [0, 1)$, for instance because the token budget stops its generation, and that a finished sample returns an answer from a distribution $\tilde{p}$ with correct mass $\tilde{p}_q$, gap $\tilde{\Delta}_q$, and pairwise quantities $\tilde{\rho}_q, \tilde{D}_q$ as in Theorem 1. Call an aggregator *supply-restricted* if it returns only an answer supplied by some finished sample and abstains when no sample finishes; truncation-aware SC, the plurality over finished samples, is one.

Proposition 1: Unfinished samples**.** *(i) Every supply-restricted aggregator abstains, and hence does not return y, with probability at least τ_q^k. (ii) If $\tilde{\Delta}_q < 0$, truncation-aware SC returns y with probability at most $\left(1 - (1 - \tau_q)\tilde{\rho}_q(1 - e^{-\tilde{D}_q})\right)^k$. (iii) If $\tilde{\Delta}_q > 0$, truncation-aware SC returns y with probability tending to one as $k \to \infty$.*

Abstention thus leaves the inversion structure intact and adds a floor τ_q^k. The floor vanishes as k grows, but slowly when chains rarely finish: pushing it below ε takes $k \geq \log\varepsilon / \log\tau_q$ samples, which for $\varepsilon = 0.01$ is seven samples at $\tau_q = 0.5$ but 90 at $\tau_q = 0.95$. At a fixed number of calls, raising the probability that a candidate finishes is therefore an alternative to adding samples; the repair candidates of Section 4 are one way to do this. Appendix D states when the number of distinct SC answers, which the decision rule of Section 4 uses as a trigger, carries information about p_q.

4 A Repair Portfolio and the Frozen Configuration

Candidates. *Self-consistency*: chain-of-thought (CoT) samples at temperature $T{=}0.7$. *Targeted repair*: a greedy CoT chain at $T{=}0$ is split into textual steps; its token log-probabilities are divided into as many equal-length token blocks as there are steps, each step receives the mean top-20 entropy of its block, the 40% of steps with the highest values are masked, and the model is asked, at $T{=}0.3$, for a corrected complete solution. The block assignment is a proxy for step-level uncertainty; the evaluation concerns the repair as a whole rather than the step-selection rule, and Appendix B gives the segmentation, entropy and prompts. *Full regeneration*: the first 6,000 characters of the greedy chain are shown with a review-and-rewrite instruction at $T{=}0.3$. *Backward cloze* (BC): from the plurality answer of three forward chains at $T{=}0.7$, a backward reconstruction at $T{=}0$ is prompted; the arm returns the plurality of the four answers. *Random-mask repair*, stored with every record, masks random steps under a different instruction and is used only as a control.

Frozen primary configuration. The configuration evaluated throughout is the *eight-call portfolio*: five SC samples, one greedy chain that is shared by the two repairs and *does not vote*, targeted repair and full regeneration, i.e. the same eight model calls as SC@8. Its decision rule is adaptive consensus-divergence routing (ACDR), defined below, which on every audited record of the original study, the archive and the prospective cohort coincides with returning the SC plurality whenever at least one of the five SC samples returns an answer, and otherwise the plurality of the returned repair answers. The comparisons are SC@8, the plurality of eight SC samples, and plain plurality over the same seven voting candidates; all $\binom{8}{5} = 56$ choices of five SC samples are averaged exactly. This configuration, its prompts, temperatures, masking fraction, thresholds and tie conventions were fixed before the archival evaluation of Section 6.2; they were chosen after the per-arm results of the original study of Section 6.1 were known. The *full portfolio* of 15 calls, eight SC samples, the greedy chain, targeted repair, regeneration and the four backward-cloze calls, is reported as a secondary configuration.

Decision rule. The released code applies ACDR, which also consults the number of distinct SC answers and an agreement threshold; Appendix B lists all branches. Under the audited evaluation it coincides with the simple rule "SC plurality if any SC sample returns an answer, otherwise the plurality of the returned repair answers, ties to targeted repair" on all 300 full-portfolio and 16,800 eight-call decisions of the original study, on all 475 and 26,600 decisions of the archive, and on every decision of the prospective cohort; under the automatic evaluation it coincides on all but 36 of the 16,800 decisions of the original study and all but 3 and 98 of the archive decisions. In the registered runs of Appendix E with a second model family and with thinking disabled, ACDR, applied as registered, differs from the simple rule on some decisions, and both are reported there. The gain is therefore attributed to the candidates, used when SC supplies no answer. *ACDR-Lazy* generates repairs only if SC returns no answer, at least four distinct answers, or a plurality below half of its returned votes; otherwise it returns the SC plurality. Appendix B proves equivalence to ACDR.

5 Evaluation Protocol

Data. The original study and the archive use Qwen3.5-27B (Qwen Team, 2026), with a 4,096-token completion budget per call and an 8,192-token context. The *original study* uses 100 MATH (Hendrycks et al., 2021) test questions from seven subjects, 53 of Level 4 and 47 of Level 5, each under seeds 11, 23 and 37, which gives 300 records. The *archive* is a 500-question MATH Level 4–5 portfolio generated earlier under the same protocol at the same budget with seed 11; all 500 questions are Algebra, 238 of Level 4 and 262 of Level 5. The *prospective cohort* of Section 6.3 uses Qwen3.5-9B (Qwen Team, 2026) under the same generation protocol on 100 MATH Level 4–5 test questions whose identifiers occur in no earlier artifact of this project, 48 of Level 4 and 52 of Level 5 stratified over the seven subjects, with seeds 11, 23 and 37 and per-call budgets of 4,096 and 16,384 tokens at contexts of 8,192 and 32,768, and the finish reason of every call is logged. Each record stores every generation, namely the eight SC samples, the greedy chain, targeted repair, full regeneration, the random-mask control repair and backward cloze, and all policies are recomputed from the stored text. All calls use the model's chat template with its default settings, under which the model writes a thinking block before its answer; a call finishes when the model ends its turn, as detailed in Appendix B.

Table 1: Audit of every scored answer. "Restored": the text states an answer that the locator missed or misparsed; "abstention": the generation stops inside its planning text without a final answer. The recovery column concerns answers newly admitted by the recovery policy of Section 6.5 in the original study and archive; for the prospective cohort at both budgets admitted answers of the mechanical rule R_U were read and only their parse was corrected, marked †, mostly by removing a closing quotation mark or full stop. One further prospective backward-cloze return was corrected to another wrong value. In parentheses: abstentions credited before the audit.

	Original	**Archive**	**Prospective**	**Recovery**
Finished non-BC candidates read	1,700	3,942	3,236	–
scored incorrect or unparsed by the locator	49	185	36	–
stated answer restored	48	167	32	–
changed to abstention	1 (0)	8 (1)	0 (0)	–
Backward-cloze returns read / made correct	300 / 15	500 / 19	600 / 21	–
Recovered read / corrected / abstention	–	–	480 / 382† / 0	281 / 36 / 71
Credited answers found to state another value	0	0	0	–

Evaluator. Two-stage candidates are stored as the greedy text followed by the repair text; only the repair text is scored, and backward cloze is scored by the answer it returned. A deterministic, gold-free locator takes the last boxed expression, else, in order, a short closing mathematical line, the value after the last answer anchor, "answer is" or "answer:", near the end, and weaker anchors listed in Appendix B; the span is parsed into a symbolic expression, and an empty parse is an abstention. Within a record, answers are clustered gold-free by symbolic equivalence in order of appearance, and votes and correctness use these classes; a class is correct if it matches the last boxed expression of the original MATH solution. SC@8 and plurality controls give fractional credit to ties; ACDR's rule is deterministic, and by Appendix B crediting its ties fractionally changes no conclusion.

Return policy. We evaluate a system that discards generations stopped by the budget: such a generation casts no vote and is scored incorrect if returned alone, even if its text already contains a candidate answer. In the prospective cohort a call is unfinished when its logged finish reason is the length limit; in the original study and the archive a call is unfinished when its completion reaches 4,096 tokens, and every SC chain that does so has exactly 4,096; a repair stage is also unfinished when its prompt plus output reaches the 8,192-token context; backward cloze, which the eight-call configuration does not use, stores its calls only as an aggregate and is scored as returned. In the original study 1,273 of 2,400 SC chains, or 53%, are unfinished and on 99 of 300 records none of the eight finishes; in the archive 1,290 of 4,000 are unfinished. Section 6.5 evaluates the alternative policy that recovers answers stated in unfinished chains.

Audit. Every answer that enters a scored comparison of the original study, the archive and the prospective cohort was read, as Table 1 summarises; the control runs of Section 6.4 are scored automatically. For every candidate that the locator scores incorrect or cannot parse, and for the 191 original-study candidates on which a decision can rest on at most three finished SC chains or on repair votes, we read the full text or its last 700 characters; for every other credited candidate we read the closing statement, and backward-cloze returns were read through parsing and class assignment. Most corrections restore a stated answer that the locator missed, because it located a variable name or intermediate expression or because a dollar sign or unit defeated the parser; a few generations stop in planning text without a final answer and become abstentions. The corrections of finished non-BC candidates fall mostly on SC samples, 42 of 49 in the original study, 154 of 175 in the archive and 20 of 32 prospectively, against 0, 4 and 8 on targeted repairs and regenerations, so the audit does not favour the repairs. The *audited* evaluation replaces each corrected located answer with the stated answer and re-clusters the record; it is primary, and the automatic evaluation is reported alongside.

Statistics. Intervals are 95% confidence intervals (CI), percentile intervals of a paired bootstrap over questions with 20,000 resamples and seed 2026, in which a question contributes the mean over its seeds and, for eight-call configurations, over its 56 SC subsets; p-values are two-sided bootstrap sign probabilities. For the archive, the registration fixed two primary endpoints, eight-call − SC@8 and full portfolio − SC@8, with Holm correction at family-wise level 0.05 on the automatic evaluation

Table 2: Original study: MATH Level 4–5, Qwen3.5-27B, 4,096-token per-call budget, 100 questions $\times$ 3 seeds. Tokens per record, each generation counted once: completion C and prompt P. Δ: paired gain over SC@8 in points with question-clustered 95% CI. Eight-call rows average all 56 five-of-eight SC subsets.

				Audited		**Automatic**	
Configuration	Calls	C	P	Acc.	Δ [95% CI]	Acc.	Δ [95% CI]
SC@8	8	27.2k	0.93k	.670	–	.667	–
Frozen eight-call	8	26.6k	5.70k	.770	+10.0 [5.8, 14.5]	.765	+9.8 [5.6, 14.4]
same 7 candidates, plurality	8	26.6k	5.70k	.770	+10.0 [5.8, 14.5]	.770	+10.3 [6.1, 14.9]
6 SC + targeted repair	8	–	–	.748	+7.8 [4.4, 11.6]	.745	+7.8 [4.4, 11.6]
6 SC + regeneration	8	–	–	.709	+3.9 [1.4, 6.7]	.707	+4.1 [1.5, 7.0]
Full portfolio, ACDR	15	50.7k	–	.777	+10.7 [7.0, 14.7]	.770	+10.3 [6.7, 14.3]
Full portfolio, lazy	8–15	36.5k	–	.777	+10.7 [7.0, 14.7]	.770	+10.3 [6.7, 14.3]

and the audited evaluation reported alongside; for the prospective cohort the two primary endpoints are eight-call $-$ SC@8 at each budget, with the same correction.

6 Results

6.1 Exploratory original study

Table 2 reports the original study. The frozen eight-call portfolio raises accuracy from 0.670 to 0.770, by $+10.0$ points $[5.8, 14.5]$ and by $+9.8$ under the automatic evaluation, with per-seed accuracies 0.774, 0.766 and 0.769, and plain plurality over the same seven voting candidates gives the same accuracy, so the candidates, not the way they are weighted, carry the gain. Replacing two SC samples by targeted repair alone adds 7.8 points and by regeneration alone 3.9 points; with targeted repair in place of the random-mask repair arm the configuration gains 8.3 points $[4.6, 12.4]$, a comparison of two repair designs with different instructions. The full portfolio adds 10.7 points, and its lazy form makes identical decisions while generating repairs on 99 of 300 records.

Cost. The eight-call portfolio uses the same number of calls as SC@8 and generates slightly fewer completion tokens, 26.6k versus 27.2k per record, with two calls rewriting an existing chain; but the repair prompts contain that chain, so it reads 5.70k prompt tokens against 0.93k, and the sum of prompt and completion tokens is 14.8% higher than SC@8's, 15.4% on H475 and 14.8% and 5.7% in the prospective cohort at the two caps. The shared greedy call is counted once; these totals count tokens, and latency is outside this cost accounting. Appendix E lists the registered comparison with fewer, longer SC chains.

6.2 Pre-registered archival evaluation

The registration fixed the configurations, the primary population H475, the two endpoints and the Holm decision rule before any current-evaluator output existed; an amendment written after the evaluator had produced its per-record outputs but before any endpoint was computed added the sensitivity populations H382 and all 500 questions, the exposure disclosure, further controls and the requirement of a complete audit. Table 3 confirms both primary endpoints: on H475 the eight-call portfolio improves on SC@8 by 3.7 points $[2.1, 5.6]$, or by $+4.4$ $[2.7, 6.4]$ under the automatic evaluation, and the full portfolio by 4.6 points $[2.7, 6.5]$, with the same direction on H382 and on all 500 questions. Plain plurality over the same seven candidates again equals the eight-call rule at $+3.7$, replacing two SC samples by targeted repair alone adds 2.7 points $[1.3, 4.3]$, and this configuration again exceeds the one with the random-mask arm in its place by $+2.5$ $[1.0, 4.1]$. The archive is a single seed of previously used Algebra questions, so we read it as a larger replication of the original direction under a fixed configuration.

Table 3: Pre-registered archival evaluation: 500-question MATH Level 4–5 Algebra portfolio, same model, generation protocol and budget, seed 11. H475: the registered primary population, the archive without the questions it shares with the original study's 100 test and 50 calibration questions; H382: H475 minus the 93 questions of an earlier evaluator-development cohort. All 500 questions were part of an earlier, legacy-evaluated SC scaling analysis. Δ over SC@8 in points, question-clustered 95% CI; Holm-adjusted decisions on the registered endpoints are all positive.

		Audited			**Automatic**		
Population	**Configuration**	SC@8	Acc.	Δ [95% CI]	SC@8	Acc.	Δ [95% CI]
H475, primary	Eight-call portfolio	.884	.922	+3.7 [2.1, 5.6]	.860	.904	+4.4 [2.7, 6.4]
	Full portfolio		.931	+4.6 [2.7, 6.5]		.916	+5.6 [3.6, 7.7]
H382	Eight-call portfolio	.882	.916	+3.3 [1.6, 5.4]	.857	.896	+3.9 [2.0, 6.0]
	Full portfolio		.921	+3.9 [2.1, 6.0]		.906	+4.8 [2.9, 7.1]
All 500	Eight-call portfolio	.880	.922	+4.2 [2.4, 6.1]	.857	.905	+4.8 [3.0, 6.8]
	Full portfolio		.930	+5.0 [3.2, 7.0]		.916	+5.9 [4.0, 8.0]

Table 4: Prospective replication, pre-registered before any generation: Qwen3.5-9B, 100 unused MATH Level 4–5 questions $\times$ 3 seeds per budget, audited evaluation. Unfinished: share of SC calls whose logged finish reason is the length limit; no SC answer: records on which none of the eight finishes. Δ over SC@8 in points, question-clustered 95% CI; Δ^{R_U}: eight-call^{R_U} $-$ SC@8^{R_U} under the mechanical recovery rule of Section 6.5.

Budget	Unfinished	No SC answer	SC@8	Eight-call	Δ [95% CI]	Δ^{R_U} [95% CI]
4,096	80%	188 / 300	.373	.467	+9.4 [5.5, 13.8]	+10.0 [5.8, 14.8]
16,384	32%	41 / 300	.863	.980	+11.7 [6.3, 17.7]	+10.1 [5.2, 15.7]

6.3 Pre-registered prospective replication on unused questions

The registration fixed the questions, drawn with a fixed seed from the 1,482 Level 4–5 test questions whose identifiers appear nowhere in this project's earlier artifacts, as well as the model, seeds, budgets, the frozen configurations, the evaluator, the audit and the two primary endpoints before any generation existed; the audit summarised in Table 1 was completed before any accuracy was computed. Table 4 confirms both primary endpoints after Holm correction: the eight-call portfolio improves on SC@8 by 9.4 points at 4,096 tokens, from 0.373 to 0.467, and by 11.7 points at 16,384 tokens, from 0.863 to 0.980, and the full portfolio and the recovery policy R_U also give positive gains with intervals above zero. The mechanism is that of the earlier studies: at 16,384 tokens a third of the SC chains still stop at the budget and all eight do on 41 of 300 records; targeted repairs and regenerations finish on 82% and 87% of calls against 68% for SC samples, and on 33% and 33% against 20% at 4,096 tokens; the eight-call portfolio is correct on 85% of the records without an SC answer, 16% at 4,096 tokens, while at 16,384 tokens both it and SC@8 are correct on all 259 records on which some SC chain finishes.

6.4 Why the gain arises: termination, not aggregation

Figure 1 decomposes the effect. Panel a shows that in both studies the repairs finish more often than fresh chains: 68% of targeted repairs and 57% of regenerations finish in the original study against 47% of SC samples and 51% of greedy chains, and 80% and 74% against 68% and 73% in the archive; the random-mask repair arm, whose instruction differs, finishes on 15% and 19%. Panel b shows that the gain arises on the records on which all SC chains are stopped. In the original study SC abstains on 99 of 300 records; the eight-call and full portfolios answer 32.3% of them correctly, and targeted repair alone is correct on 25.3%, regeneration on 15.2% and the greedy chain the repairs start from on 5.1%. In the archive SC abstains on 54 of 475 records, or 11%; the eight-call portfolio answers 35.2% of them and the full portfolio 40.7%. On the remaining records SC@8 and the full portfolio are correct on 100% in the original study, 200 of 201 under the automatic evaluation, and on 99.8% in the archive, and return the same answers; the eight-call portfolio, which keeps five of the eight

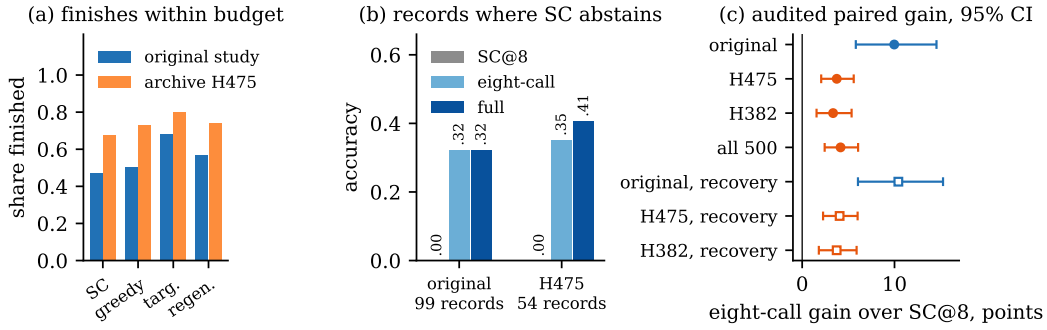


Figure 1: Both studies, audited evaluation. (a) Share of candidates that finish within the budget, and for repairs also within the context, by type. (b) Accuracy on the records on which no SC chain returns an answer and SC@8 abstains. (c) Paired gain of the eight-call portfolio over SC@8 with question-clustered 95% CIs, shown as circles under the declared abstention policy and as squares under the registered recovery policy of Section 6.5.

SC samples, is correct on 99.0% and 99.5% of them, so its net gain is the supply gain less the effect of voting with five instead of eight SC samples. The gain thus tracks the rate at which SC abstains, which explains the smaller effect in the archive, where SC chains finish more often.

More SC samples at the same cap. Pooling the SC samples of the three seeds of each question gives SC@16 and SC@24 at the same per-call cap; this analysis was registered before any of its numbers was computed. The full portfolio, with 15 calls and 50.7k completion tokens per record in the original study, exceeds SC@16, with 16 calls and 54.4k, by 5.7 points $[3.0, 8.7]$ in the original study and by 9.3 $[5.0, 14.3]$ and 8.3 points $[3.7, 13.7]$ at the two prospective caps, all three confirmed after Holm correction, with backward cloze scored as returned; the eight-call portfolio, which uses no backward cloze, exceeds SC@16 by 5.0 points $[1.6, 8.7]$ in the original study, and at 16,384 tokens it exceeds SC@24 by 9.0 points $[3.7, 15.0]$ with a third of the calls; Appendix E lists every comparison. Counting prompt and completion tokens, these two comparisons give the eight-call portfolio 57% of SC@16's tokens, 32.3k against 56.3k per question, and 35% of SC@24's, 87.5k against 248.2k. Extra samples close the abstention floor as slowly as Proposition 1 predicts: the share of questions with no SC answer falls only from 0.32 at $k = 8$ to 0.25 at $k = 24$ in the original study and from 0.61 to 0.54 at 4,096 tokens prospectively, close to $\mathbb{E}_q\, \tau_q^k$ from each question's pooled no-answer rate, 0.33, 0.28, 0.63 and 0.57, a descriptive check on the same counts.

Why do the repairs finish? Figure 2 in Appendix A shows shorter repair stages, a median of about 3.0k tokens for targeted repair against the full budget for SC samples; these lengths and the example there are consistent with repairs that, conditioned on an existing attempt, write out a solution instead of re-deriving and re-checking it.

What differs between the repairs and fresh chains. The repairs differ from SC samples in temperature, 0.3 against 0.7, in conditioning on a prior attempt, and, in the original study and the archive, in an instruction to end with an answer line; in the prospective cohort the SC samples carry that instruction too, and Appendix B gives all prompts. The greedy chain, sampled at temperature zero from the SC prompt, finishes about as often as SC samples, 51% against 47% in the original study and 73% against 68% in the archive, and is correct on 5.1% and 5.6% of the records without an SC answer, against 25.3% and 25.9% for targeted repair. Two registered controls regenerate the SC samples of the prospective cohort at 4,096 tokens with all else fixed, scored by the automatic evaluation and listed in Appendix E. Drawn at the repairs' temperature 0.3 with their final-line instruction, SC samples finish on 22% of calls against 20% at 0.7, and the eight-call portfolio exceeds their SC@8 of 0.378 by 8.7 points $[4.7, 13.2]$, confirmed after Holm correction, and by 9.4 points under R_O: matching the repairs' temperature and final-line instruction leaves the advantage in place. A budget instruction appended to the SC prompt, "Your entire response, including any thinking, must fit within 4,096 tokens…", is a second way to change the candidates: it doubles the share of SC calls that finish and raises SC@8 by 20.5 points $[14.0, 27.3]$, and the repairs add to it: in a separately

Table 5: Recovery policy R, registered before any recovery number: an unfinished candidate is admitted if its capped text contains an explicit final statement, the last boxed value or else the last strong answer anchor near the end; all other rules unchanged. Audited evaluation; the 281 newly admitted answers were read, 107 corrected, 71 of them to abstentions. Δ over SC@8^R in points, question-clustered 95% CI.

Population	SC@8^R	Eight-call^R	Δ eight-call^R [95% CI]	Δ full^R [95% CI]
Original study, 100 q $\times$ 3	.700	.804	$+10.4$ [6.0, 15.3]	$+9.0$ [5.7, 12.7]
Archive H475	.899	.940	$+4.1$ [2.3, 6.0]	$+4.4$ [2.7, 6.3]
Archive H382	.898	.935	$+3.7$ [1.8, 5.9]	$+3.9$ [2.1, 6.0]

registered run with the instruction in every prompt, the eight-call portfolio exceeds SC@8 of the same run by 9.7 points $[5.1, 14.7]$, from 0.580 to 0.677, confirmed, with fewer completion tokens, 28.1k against 28.4k per record. A third registered control regenerates every prospective targeted repair from its logged greedy chain next to two prompt-matched variants, the same prompt with as many steps marked at random and an unmasked rewrite with the matched instruction, listed in Appendix E. At 16,384 tokens the targeted-repair prompt returns a correct answer on 88% of records against 76% for the unmasked rewrite, $+12.7$ points $[8.0, 17.7]$, confirmed after Holm correction, and $+13.0$ audited; this contrast concerns marked against unmarked rewriting, not the choice of the marked steps.

A descriptive lazy eight-call variant. Because on the audited records of the three main evaluations the decision uses the repairs only when no SC sample answers, the three extra calls can there be deferred: generate the five SC samples, and the greedy chain and its two repairs only if none of them returns an answer. This makes exactly the eight-call decisions, with accuracy 0.770 in the original study and 0.922 on H475, using 6.1 and 5.4 calls per record and a prompt-plus-completion total 15% and 28% below SC@8's; Appendix A gives the token counts. Where SC abstains, deferral adds two sequential rounds, so the saving is measured in tokens; the variant was derived afterwards.

6.5 Answer recovery instead of abstention

A deployment could instead read answers already stated in stopped chains; we registered a gold-free recovery policy before computing it and applied it to every candidate, SC samples and repairs alike. Table 5 shows that recovery raises SC@8 by 3.0 points in the original study and 1.5 points on H475, and the eight-call advantage remains at $+10.4$ and $+4.1$; many stopped chains state an answer only in planning text, and the audit made 71 of the 281 admitted answers abstentions, such as "the final answer is clearly stated". Two later registered rules, detailed in Appendix A, are applied by code to every unfinished candidate: R_U, with only the parse of the located value checked, and R_O, without any manual step, which takes the value of each answer anchor's own statement and applies one fixed canonicalisation to all candidates: the eight-call advantage is then $+10.7$ $[6.3, 15.6]$ in the original study, $+3.9$ $[2.2, 5.9]$ on H475, $+3.6$ $[1.7, 5.7]$ on H382, and $+10.3$ $[6.1, 15.1]$ and $+10.1$ $[5.2, 15.7]$ at the two prospective caps. The advantage is thus not an artifact of the abstention policy.

7 Discussion and Limitations

Under a hard per-call cap, replacing three of eight SC samples by a greedy chain and two repairs improves accuracy at equal call count: about ten points in the exploratory study, four in the pre-registered archive and nine to twelve prospectively, including when answers are recovered from stopped chains. The repairs finish more often and answer where SC does not; on the audited records of the three main evaluations the rule reduces to "use SC if it answers, else the repairs", so the repairs can be deferred. The scope is Qwen3.5 in its given thinking mode on MATH Level 4–5, at caps that many SC chains exhaust: a third do so even at 16,384 tokens for the 9B model. Interventions inside a call, such as budget forcing, act on single chains and combine with any choice of candidates; spending the three calls on continuing stopped SC chains is a further allocation outside this evaluation. The gain equals the SC no-answer rate times portfolio accuracy on those records, less the effect of voting with five instead of eight SC samples elsewhere; both are estimable from a sample at the given cap. Appendix E reports registered boundary cases in which SC chains mostly finish.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to generate synthetic data sets, and surveys, interviews and the transcription of recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies inference-time procedures for openly released language models on the public mathematics benchmarks MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021). It involves no human subjects, no crowdsourcing and no personal data: the only data are benchmark questions with their reference answers and the text generated by the evaluated models, which the supplementary material releases in scored form together with the closing text of the candidates. The procedures change only how many and which candidate answers a model produces within a fixed number of calls; they do not add capabilities beyond those of the underlying model, and we see no specific harmful or dual-use application beyond those of the models themselves. Accuracy gains on mathematics questions do not certify the correctness of individual answers, and the results should not be read as evidence of reliability in high-stakes use. All results are reported with the evaluation, audit and registration procedures that produced them, and every primary registered endpoint of the studies reported here is listed in Appendix E, with the secondary analyses of these runs in the supplementary material.

REPRODUCIBILITY STATEMENT

The candidate generators, the frozen eight-call configuration and the decision rule are specified in Section 4; the exact prompts, sampling settings, step segmentation and entropy computation are given in Appendix B. The data, return policy, evaluator, audit and statistical procedures are described in Section 5, and the recovery rules in Section 6.5 and Appendix A. The modelling assumptions and complete proofs are in Section 3 and Appendices C and D. Every primary registered endpoint of the studies reported here is listed in Appendix E; the supplementary material prints the secondary analyses of these runs. The anonymous supplementary material contains the scored candidate records of all cohorts including GSM8K, the same-cap SC controls and the registered runs of Table 12, every change made by the audit and the GSM8K audit adjudications, the closing text of every candidate other than backward cloze in the original study, the archive and the prospective cohort, the recovery admissions and their change record, the prompt templates with serialised model inputs, and a summary of the registered analysis plans; a set of scripts recomputes the policy decisions, accuracies and bootstrap intervals of the tables without any model call, apart from the recovery rows of the same-cap controls, and another script rebuilds the located answers, answer classes and correctness from the closing texts and the reference answers without reading the stored labels.

REFERENCES

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

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45767–45790, 2024. doi: 10.52202/079017-1455.

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 long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

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

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785. URL https://archive.org/details/essaisurlapplic00conggoog.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in LLMs. *arXiv preprint arXiv:2510.17472*, 2025.

Austin Feng, Marius Alonso, Ambroise Odonnat, Vasilii Feofanov, and Ievgen Redko. Optimal self-consistency for efficient reasoning with large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1274.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Wenyang Hu, Gregory Kang Ruey Lau, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35558–35572. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1801. URL https://aclanthology.org/2025.emnlp-main.1801/.

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

Siyuan Huang, Zhiyuan Ma, Jintao Du, Changhua Meng, Weiqiang Wang, and Zhouhan Lin. Mirror-Consistency: Harnessing inconsistency in majority voting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2408–2420. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-emnlp.135.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

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

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. Plan, verify and switch: Integrated reasoning with diverse X-of-Thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2807–2822.

Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.169. URL https://aclanthology.org/2023.emnlp-main.169/.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019.

Lin Mu, Wenhao Zhang, Yiwen Zhang, and Peiquan Jin. DDPrompt: Differential diversity prompting in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 168–174. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.17. URL https://aclanthology.org/2024.acl-short.17/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6023–6029. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.349.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. 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, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800.

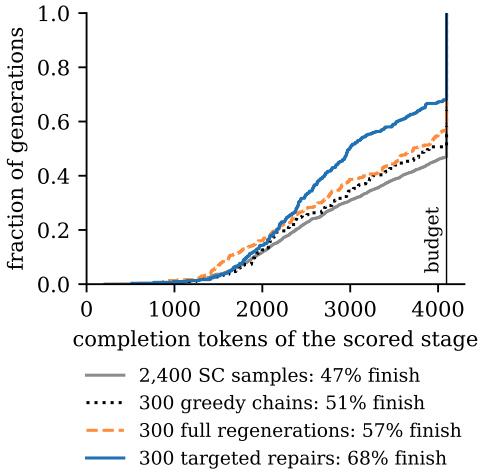
Yingchuan Zhang, Terry Ma, Wenxuan Zhong, and Ping Ma. A single revision step improves token-efficient LLM reasoning. *arXiv preprint arXiv:2602.02828*, 2026.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*, 2023.

A Additional Results

Difficulty. Consistent with the mechanism, the gain is largest where SC is weakest and abstains most: on Level 5 questions the eight-call portfolio adds 13.6 points in the original study and 5.7 in the archive, against 6.8 and 1.6 on Level 4, as Table 6 shows.

Original study. The full-portfolio gain is positive on each seed separately, at $+11$, $+14$ and $+7$ points, each with a per-seed interval excluding zero, and the eight-call accuracies per seed are 0.774, 0.766 and 0.769. Of the 32 records without an answering SC chain that the full portfolio answers correctly, 27 rest on a finished targeted repair or regeneration and five on backward cloze alone; treating backward-cloze answers whose four calls all reached the budget as unfinished gives the full portfolio 0.780, a gain of $+11.0$.



MATH record 3931, seed 11, reference answer -89. *SC samples:* all eight reach the 4,096-token budget; each has already derived $M + N = -89$ and keeps re-checking the configuration, asking for instance "Is it possible the question implies $0 < x < 180$?", until the budget stops it.

Greedy chain: also stopped while re-checking.

Targeted repair of the greedy chain: a complete solution in 2,186 tokens ending "$M + N = 1 + (-90) = -89$. The answer is -89."

Figure 2: Original study. Left: completion tokens of the scored stage for each candidate type, with the share that finishes; repairs stopped by the context limit are placed at the budget. Right: a record on which no SC chain finishes and targeted repair finishes with the correct answer.

Table 6: Gains by MATH level: paired gain over SC@8 in points with question-clustered 95% CI under the audited evaluation; the automatic point estimate follows in parentheses.

Study	**Level**	Questions	SC@8	Full portfolio	Eight-call portfolio
Original	Level 4	53	.736	+8.2 [3.8, 13.2] (+7.5)	+6.8 [2.7, 11.5] (+6.6)
	Level 5	47	.596	+13.5 [7.8, 19.9] (+13.5)	+13.6 [6.4, 21.9] (+13.4)
Archive H475	Level 4	224	.969	+2.2 [0.4, 4.5] (+4.5)	+1.6 [0.0, 3.5] (+3.4)
	Level 5	251	.809	+6.8 [4.0, 10.0] (+6.6)	+5.7 [2.9, 8.8] (+5.4)

Archive. On H475 the full portfolio's lazy form invokes the repairs on 54 records, and prompt tokens per record are 0.78k for SC@8, 4.98k for the eight-call and 5.77k for the full portfolio, with 24.1k and 23.8k completion tokens for SC@8 and the eight-call portfolio. On the 54 records without an SC answer, targeted repair is correct on 25.9%, regeneration on 18.5%, backward cloze on 16.7% and the greedy chain on 5.6%.

Token accounting. SC and greedy candidate counts refer to one call; each repair candidate's stored prompt and completion counts instead include its greedy parent plus its repair stage. Backward cloze includes all four calls. For either token type, let C_S sum the five selected SC calls, G be the greedy count, and T, F the cumulative targeted-repair and regeneration counts. The eight-call total is $C_S + T + F - G$, which counts the shared greedy call once. Lazy execution pays C_S and adds $T + F - G$ only when repairs are needed. Applying this accounting to the recorded fields reproduces all reported totals. These are token-accounting results; latency and floating-point operation counts are outside this accounting.

Lazy eight-call variant. Per record, the lazy variant of Section 6.4 generates 21.2k completion and 2.8k prompt tokens in the original study against 27.2k and 0.9k for SC@8, and 16.6k and 1.3k on H475 against 24.1k and 0.8k.

Mechanical recovery rule R_U. Amendment R-A1 to the recovery registration was written after the audited results of Table 5 were known and before R_U was computed. For every candidate that is unfinished under the audited policy, R_U takes the content of the last boxed expression of the stage text if the box is closed and its content is value-bearing, that is, a digit, a LaTeX command or mathematical symbol, a short symbolic expression, or an English value of at most two words outside a fixed list of non-answer words such as "best"; otherwise it takes the value of the last strong answer anchor starting

in the last 800 characters whose value is value-bearing and not cut off by the budget. An anchor followed by a placeholder such as "The answer is <your answer>." is skipped, and nothing else admits an answer. In particular, no value is taken from the line after such a placeholder; a complete final statement ending the text, such as "The answer is 1.", is not cut off; format placeholders such as "[number]" and a lone closing quotation mark are not values; and when an anchor's statement runs into a following placeholder sentence, the anchor's own value is used. This fallback has a narrow form: the anchor's own statement is used only when the value of the full sentence is a placeholder, so a complete anchor whose full-sentence value is empty or a non-answer word, such as "The answer is $\frac{1}{4}$." followed by planning text, stays unadmitted under R_U; rule R_O below uses the anchor's own statement for every anchor. R_U admits 237 answers, 97 in the original study and 140 in the archive, of which 19 come from boxes. The only audit read every admission whose value or admission differs from the audited answers under R and corrected the parse of the located value against its own anchor sentence, with 44 corrections such as "definitely 4." to 4; it cannot admit or remove a candidate, so a planning sentence such as "the answer is useless" stays admitted as a wrong answer. Relative to audited R, 200 admissions have the same value, 58 candidates are abstentions under both, 21 are admitted only by R_U, 13 are admitted by R_U where the audit abstained, 7 only by the audit, and 3 differ in value. SC@8^{R_U} and eight-call^{R_U} are 0.698 and 0.806 in the original study, 0.895 and 0.935 on H475 and 0.893 and 0.929 on H382; the eight-call advantage is $+10.8$ $[6.3, 15.8]$, $+4.0$ $[2.2, 5.9]$ and $+3.7$ $[1.8, 5.8]$ points, and the full-portfolio advantage $+9.5$ $[6.0, 13.3]$, $+4.0$ $[2.3, 5.9]$ and $+3.7$ $[1.8, 5.8]$ points.

Fully mechanical recovery rule R_O. Amendment R-A3, written after all R and R_U results were known and before any R_O output existed, specifies a rule without any manual step. For every unfinished candidate, SC samples and repairs alike, the box branch is that of R_U; for every strong anchor in the last 800 characters, the statement is the anchor's own statement, cut at the first line break, closing quotation mark or backtick, or at a full stop followed by white space, and the placeholder, value-bearing, cut and balance tests of R_U are applied to it; the last anchor that passes gives the answer. Every admitted anchor value is then canonicalised by one fixed function that writes the principle of the R_U parse audit as code: it cuts at a closing quotation mark, backtick or line break and drops a leading "definitely", "exactly", "robustly" or "likely", a trailing "or similar", the text of a following sentence, a trailing full stop and enclosing dollar signs. R_O admits 249 answers in the original study and the archive, 101 and 148 respectively with 19 from boxes, and 509 in the prospective cohort, 342 at 4,096 and 167 at 16,384 tokens. Against the final R_U admissions after its parse audit, 13 and 32 candidates are admitted only by R_O, 1 and 3 only by R_U, and 3 and 13 have a different located value; the complete change record is released. The candidates admitted only by R_O include complete anchors that the narrow form of R_U leaves out, for example "The answer is $\frac{1}{4}$." followed by planning text, admitted as $\frac{1}{4}$, a final line "The answer is 4." followed by a quoted format instruction, admitted as 4, and $\frac{24}{5}$, 340 and 40 in the same situations. SC@8^{R_O} and eight-call^{R_O} are 0.703 and 0.810 in the original study, 0.897 and 0.936 on H475, 0.895 and 0.931 on H382, and, in the prospective cohort, 0.503 and 0.606 at 4,096 and 0.873 and 0.974 at 16,384 tokens. The full-portfolio advantage under R_O is $+9.3$ $[6.0, 13.0]$, $+4.2$ $[2.5, 6.1]$ and $+3.9$ $[2.1, 6.0]$ points in the original study, H475 and H382, and $+11.1$ $[7.5, 14.9]$ and $+9.0$ $[4.3, 14.3]$ points at the two prospective caps.

Prospective cohort. The audited accuracies of SC@8, the eight-call and the full portfolio are 0.373, 0.467 and 0.527 at 4,096 tokens and 0.863, 0.980 and 0.970 at 16,384 tokens, with automatic full-portfolio gains of $+13.0$ $[9.0, 17.3]$ and $+10.2$ $[5.3, 15.7]$; the automatic eight-call gains are $+9.2$ $[5.3, 13.7]$ and $+11.5$ $[6.2, 17.5]$, and scoring unfinished calls by the length proxy instead of the logged finish reason changes no estimate. Plain plurality over the same seven candidates gives $+9.4$ and $+11.8$ points. Greedy chains finish on 21% and 58% of calls. Per record the eight-call portfolio generates 30.0k and 75.2k completion tokens against 30.4k and 81.7k for SC@8 and reads 6.1k and 12.3k prompt tokens against 1.0k. With R_U, SC@8^{R_U} and eight-call^{R_U} are 0.492 and 0.592 at 4,096 tokens and 0.873 and 0.974 at 16,384 tokens, and the full-portfolio advantage is $+11.2$ $[7.7, 15.0]$ and $+9.0$ $[4.3, 14.3]$ points; without the parse audit the eight-call advantages are $+9.7$ $[5.3, 14.5]$ and $+10.1$ $[5.2, 15.7]$. R_U admits 480 answers in this cohort, 315 at 4,096 and 165 at 16,384 tokens. Under evaluator F, which differs from the evaluator of Section 5 only in its answer locator and which the cohort's original registration froze as a registered sensitivity row, the eight-call gains are $+10.8$ $[6.7, 15.3]$ and $+10.9$ $[5.9, 16.6]$ points.

B Decision Rule, Evaluator and Audit Details

Prompts and sampling. Every prompt is placed in the model's chat template as the query turn, after the system message "You are a helpful assistant. /no_think". The template is applied with its default settings; this model family does not interpret "/no_think" in a message as a switch, so the serialised input ends with the opening of the assistant turn followed by the opening tag of a thinking block, every call first writes its reasoning inside that block, and the stored text of a call contains the thinking block and the answer after it. In the prompts below, italics mark inserted text and a slash marks a line break. In the original study and the archive, SC samples and the greedy chain use "Question: *question* / Let's think step by step."; in the prospective cohort they use "Question: *question* / Show the key steps, then end with exactly one final line: The answer is …", the final-line instruction of the repairs. The eight SC samples are drawn at temperature 0.7 and the greedy chain at temperature 0; all calls use top-p 1 without top-k truncation and stop only at the model's end-of-turn tokens or at the cap. Both repairs are sampled at temperature 0.3 from the following prompts, given in the prospective cohort's wording; in the original study and the archive every repair prompt ends with "End with: The answer is <your answer>." instead of the final-line instruction shown.

- *Targeted repair:* "Question: *question* / A previous solution attempt found some uncertain steps. The steps marked with [HIGH UNCERTAINTY] likely contain errors. / Step 1: *step 1* / … / Step j: [HIGH UNCERTAINTY (entropy=*value*) - NEEDS CORRECTION] / … / Rewrite the complete solution, correcting the uncertain steps. End with exactly one final line: The answer is …"
- *Full regeneration:* "Question: *question* / A student wrote this solution: / *first 6,000 characters of the greedy chain* / Review for errors and write a corrected, complete solution. End with exactly one final line: The answer is …"
- *Random-mask control repair:* as targeted repair, with the header "A previous solution has some steps masked. Fill in the missing steps:", masked steps shown as "[MASKED - fill in this step]", and the closing instruction "Complete the solution by filling in all masked steps." followed by the same final-line instruction.
- *Backward cloze:* three chains from the SC prompt at temperature 0.7; their plurality answer a is inserted in a prompt that asks the model, at temperature 0, to reconstruct the reasoning backward from "Final result: a" to the premises, to state whether a is correct and otherwise to give the correct answer, with the same final-line instruction.

Step segmentation and entropy. The greedy text is split at line breaks followed by an optional "Step" and a number with ".", ")" or ":" when this yields at least three parts, otherwise at sentence ends followed by a capital letter, otherwise at line breaks; parts with fewer than three words are dropped. The entropy of a generated token is that of the renormalised top-20 next-token distribution. The N generated tokens are divided into as many consecutive blocks of $\lfloor N/n \rfloor$ tokens as there are steps, n, each step receives the mean entropy of its block, and the $\max(1, \lfloor 0.4n \rfloor)$ steps with the highest values are masked; the random-mask control draws the same number of steps from a generator seeded by the question and the run seed.

Executed ACDR rule. Given the SC answers actually returned, ACDR computes their number of distinct answers n_{dist}. (i) If no SC sample returns an answer, it outputs the plurality of the repair answers, ties broken in the order targeted repair, backward cloze, full regeneration; if no repair returns an answer either, it abstains. (ii) If $n_{\mathrm{dist}} < 4$, it outputs the SC plurality, with ties going to the first-appearing answer, when that plurality holds at least half of the SC votes; otherwise it outputs the repair plurality if at least a fraction 0.67 of the returned repair answers agree, and the SC plurality if they do not or if no repair returns an answer. (iii) If $n_{\mathrm{dist}} \geq 4$, it scores each answer a by

$$s(a) = (1 - w)\,\frac{c_a}{k'} + w\,\frac{r_a}{m'}, \qquad w = \sigma(\alpha H_{\mathrm{SC}}), \quad \alpha = 2,$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function, H_{SC} is the natural-logarithm entropy of the empirical distribution of the returned SC answers, c_a, r_a are SC and repair votes and k', m' the numbers of returned SC and repair answers; when no repair returns an answer, $m' = 0$ and the repair term is zero, so the score orders answers by SC votes. ACDR returns the answer with the highest score; among equal scores the SC plurality comes first and the other answers follow the lexicographic order of their class representatives. With $k' \leq 8$ and $n_{\mathrm{dist}} \geq 4$, $H_{\mathrm{SC}} \geq 1.0735$, so one additional repair vote always outweighs any difference in SC share and the score orders answers by

repair votes, then SC votes, for every $\alpha \geq 1$. With three returned repair answers the threshold 0.67 lets the repair plurality override the SC plurality only when all three agree, and with the two repair answers of the eight-call portfolio only when both agree. ACDR-Lazy requests repairs if $k' = 0$, $n_{\text{dist}} \geq 4$, or $\max_a c_a < k'/2$. Otherwise branch (ii) returns the SC plurality regardless of any repair answers, so omitting repairs is exact; in every other case the complete rule is evaluated. This covers the low-plurality branch even when only three distinct SC answers occur. Under the audited original, archival and prospective evaluations, the low-plurality condition never fires, so on these cohorts the reported lazy decisions, trigger counts and token costs are those of the trigger without this condition; the exhaustive check below covers every other released cohort. The supplementary checks exhaust all returned-vote count profiles up to eight SC votes and three repairs and re-evaluate every released set of scored records; the lazy and complete rules agree throughout. The constants $n^*_{\text{dist}} = 4$ and $\alpha = 2$ are fixed design constants; by the argument above, the executed order is the same for every $\alpha \geq 1$.

Ties. When no SC answer is returned, ACDR breaks repair-plurality ties deterministically. In the original study this happens on 19 of the 99 such records; crediting those ties fractionally gives the full portfolio 0.762, a gain of $+9.2$ $[6.2, 12.5]$, and 0.753 with a gain of $+8.7$ under the automatic evaluation. On the archive H475, fractional crediting gives 0.925, a gain of $+4.1$ $[2.5, 5.8]$. For the eight-call portfolio, crediting ties between its two repair answers fractionally changes no decision in the original study, on H475 or at 4,096 prospective tokens, and one record at 16,384 tokens, which gives 0.982 and $+11.8$ $[6.5, 17.8]$.

Locator. After the last boxed expression, the locator tries a final line of at most 40 characters consisting only of a mathematical expression; the value in the sentence after the last "answer" anchor followed by "is", ":" or "=" within the last 800 characters; then a bold value in the final line, a closing display equation, the value after the last weak anchor, the last mathematical span of the final paragraph and the last sentence. The span is canonicalised by removing a trailing period and an enclosing text-mode wrapper and, for an equation chain without inequalities, by keeping the part after the last "=". The evaluator keeps structural commas in references, which occur in 7 of the 100 original-study references, scores two-stage candidates on their repair text alone and votes over answer classes rather than raw strings.

Audit material. The supplementary material contains every scored record before and after the audit with the finish status, located answer, answer class and tokens of each candidate, a list of every change the audit made, the closing 1,200 characters and last box value of every candidate other than backward cloze in the original study, the archive and the prospective cohort, the recovery admissions, a script that recomputes the policy decisions, accuracies and intervals of the tables from these records, and a script that rebuilds every located answer, answer class and correctness label from the closing texts, the audit list and the reference answers; the rebuilt labels agree with the released ones on all 18,200 MATH candidates. In the archive, one question, MATH record 333, has a reference that differs from the answer all eleven of its finished candidates state; we keep the reference.

C Proofs for Section 3

Answer-only rules. The bound is not an artifact of plurality counting. Call an aggregator *co-plurality-restricted* if it maps the multiset of answers to one of the answers with the largest empirical count, with any label-free tie rule, weighting or confidence threshold on top.

Proposition 2: Answer-only barrier**.** *For an inverted question with $\varepsilon_q = p^*_q - p_q > 0$, every co-plurality-restricted aggregator returns y with probability at most $\left(1 - \rho_q + \rho_q e^{-\varepsilon_q^2/(2\rho_q^2)}\right)^k$.*

The class is narrow by design: verifiers, learned rerankers and language-model judges use information beyond answer counts and are not covered. Within this class, the probability of returning y on an inverted question tends to zero as samples from the same process are added, even though a few extra samples can raise it at small k; candidates with a different answer distribution are one way out, and information beyond the answers is another.

Proof of Theorem 1. Part 1. By the strong law of large numbers, for each q the empirical frequency of every answer converges almost surely to its probability. If $\Delta_q > 0$, y eventually has strictly

more votes than every wrong answer, so SC returns y for all large k; if $\Delta_q < 0$, some wrong answer eventually has strictly more votes than y, so SC does not return y for all large k. Since $\Pr_D(\Delta_q = 0) = 0$, the indicator that SC returns y converges to $\mathbf{1}[\Delta_q > 0]$ for D-almost every q, and dominated convergence gives $A(k) \to \Pr_D(\Delta_q > 0) = 1 - r_P$. *Part 2.* Since $A(k) \to 1 - r_P < A(k_0)$, there is $K > k_0$ with $A(k) < A(k_0)$ for all $k \geq K$, so A is not monotone non-decreasing, and $\sup_k A(k) = \max_{k<K} A(k)$ is attained at a finite k^*. *Part 3.* Let $m = n_y + n_{a^*}$. Then $m \sim \text{Binomial}(k, \rho_q)$ and $n_y \mid m \sim \text{Binomial}(m, \pi_q)$ with $\pi_q < \frac{1}{2}$, and the Chernoff bound gives $\Pr(n_y \geq m/2 \mid m) \leq e^{-mD_q}$; on the event $m = 0$ both sides equal one, so the bound holds for every m. Taking expectations with the binomial moment generating function, $\Pr(n_y \geq n_{a^*}) \leq \mathbb{E}[e^{-mD_q}] = (1 - \rho_q + \rho_q e^{-D_q})^k$, and $-\log(1-x) \geq x$ gives the lower bound on the rate. The conditional rate $\rho_q D_q$ obtained by substituting $m \approx \rho_q k$ is not a valid rate for this unconditional probability. □

Proof of Proposition 2. Such a rule returns y only if n_y is maximal, hence only if $n_y \geq n_{a^*}$. Conditionally on m, Hoeffding's inequality gives $\Pr(n_y/m - \pi_q \geq \frac{1}{2} - \pi_q \mid m) \leq e^{-2m(1/2-\pi_q)^2} = e^{-m\varepsilon_q^2/(2\rho_q^2)}$, since $\frac{1}{2} - \pi_q = \varepsilon_q/(2\rho_q)$; on the event $m = 0$ the ratio n_y/m is undefined, but the conditional probability is at most one, which equals the right-hand side at $m = 0$. Averaging over $m \sim \text{Binomial}(k, \rho_q)$ gives the bound. □

Proof of Proposition 1. (i) With probability τ_q^k no sample finishes. (ii) The number of samples that finish and equal y or the modal wrong answer is $\text{Binomial}(k, (1-\tau_q)\tilde{\rho}_q)$, and the argument of Theorem 1, part 3, applies with this binomial. (iii) Since $\tau_q < 1$, the number of finished samples tends to infinity almost surely, and they are independent draws from $\tilde{p}$, so part 1 applies. □

D Answer Diversity as a Signal of Low Correct Mass

A natural label-free signal is n_{dist}, the number of distinct answers among the k samples, a monotone transform of the cluster count that Wang et al. (2024) use as a confidence measure. We ask when it carries information about p_q. Under a uniform wrong-answer model, $\Pr(X = a) = (1-p)/(C-1)$ for each wrong answer,

$$\mathbb{E}[n_{\text{dist}} \mid p] = 1 - (1-p)^k + (C-1)\Big[1 - \Big(1 - \tfrac{1-p}{C-1}\Big)^k\Big].$$

Theorem 2: n_{dist} discriminability. *Let* $G(\varepsilon) = \mathbb{E}[n_{\text{dist}} \mid p = \frac{1}{2} - \varepsilon] - \mathbb{E}[n_{\text{dist}} \mid p = \frac{1}{2} + \varepsilon]$. *Then*

$$G(\varepsilon) = 2\varepsilon k\Big[\Big(1 - \tfrac{1}{2(C-1)}\Big)^{k-1} - \big(\tfrac{1}{2}\big)^{k-1}\Big] + O(\varepsilon^2).$$

(i) For $C \geq 3$ and $k \geq 2$ the coefficient is strictly positive, so n_{dist} is locally larger below $p = 0.5$ than above it; more generally $\frac{d}{dp}\mathbb{E}[n_{\text{dist}} \mid p] < 0$ for every $p > 1/C$. (ii) For $C = 2$, $\mathbb{E}[n_{\text{dist}} \mid p] = 2 - p^k - (1-p)^k$ is symmetric about $p = 0.5$: n_{dist} cannot distinguish p from $1-p$, and $G \equiv 0$. (iii) The leading coefficient increases with C; for fixed C it is maximised at an intermediate k.

Proof. With $w = (1-p)/(C-1)$, $\frac{d}{dp}\mathbb{E}[n_{\text{dist}} \mid p] = k(1-p)^{k-1} - k(1-w)^{k-1}$, which is negative iff $w < p$, i.e., iff $p > 1/C$. Since $G'(0) = -2\frac{d}{dp}\mathbb{E}[n_{\text{dist}} \mid p]\big|_{p=1/2}$ equals the displayed coefficient, which is positive for $C \geq 3$ and $k \geq 2$, continuity gives $G(\varepsilon) > 0$ for small $\varepsilon > 0$. For $C = 2$ the expectation is symmetric about $\frac{1}{2}$. The coefficient increases in C because $1 - \frac{1}{2(C-1)}$ does, and for fixed C the term $k\eta^{k-1}$ with $\eta < 1$ is unimodal in k. □

The theorem concerns the binary boundary $p = 0.5$, which is a necessary envelope for inversion rather than the inversion boundary itself; it is a local statement under a uniform wrong-answer model, and concentrated wrong answers lower n_{dist} for any fixed p. Its practical content is a design rule: for open-ended answers, high diversity points to low correct-answer mass; for binary answers, n_{dist} depends on p only through $|p - \frac{1}{2}|$, so it can flag uncertainty but cannot tell whether the modal answer is the correct one. ACDR of Section 4 uses n_{dist} in exactly this role, as a trigger for changing the generation process, not as a calibrated estimate of Δ_q.

E REGISTERED TESTS

Registered confirmatory families. Table 7 lists every endpoint of the three registered confirmatory families as registered, on the automatic evaluation and the primary populations with Holm correction over the two endpoints of each family, Table 8 the remaining registered comparisons of the prospective cohort, Table 9 every comparison of the registered pooled-SC analysis, Table 10 every comparison of the registered same-cap SC controls, Table 11 every comparison of the registered run with a budget instruction in every prompt, and Table 12 the registered runs with a second model family, with thinking disabled, and with prompt-matched repair variants.

Table 7: Registered confirmatory endpoints, automatic evaluation, as registered. Δ in points with question-clustered 95% CI; p: two-sided bootstrap sign probability; Holm at family-wise level 0.05.

Registration: population	**Endpoint**	Δ	95% CI	p	Holm
Archive H500: H475	full portfolio − SC@8	$+5.58$	[3.58, 7.68]	< 0.001	reject
	eight-call − SC@8	$+4.45$	[2.67, 6.41]	< 0.001	reject
GSM8K G1319: 1,169 q	eight-call − SC@8	$+0.11$	$[-0.26, 0.54]$	0.62	retain
	full portfolio − SC@8	$+0.43$	[0.13, 0.81]	0.009	reject
Prospective: 100 q × 3	eight-call − SC@8, 4,096	$+9.23$	[5.27, 13.69]	< 0.001	reject
	eight-call − SC@8, 16,384	$+11.52$	[6.21, 17.49]	< 0.001	reject

Table 8: Remaining registered comparisons of the prospective cohort, secondary and without a decision rule, audited evaluation; E4 counts records with at least one finished SC answer under evaluator F; the evaluator-F sensitivity row is in Appendix A. Δ in points with question-clustered 95% CI; token counts per record as completion / prompt.

Comparison	4,096	16,384
eight-call − SC@8^{R_U}	-2.4 [-7.5, 2.7]	$+10.7$ [5.7, 16.3]
SC@8^{R_U} − SC@8	$+11.8$ [7.8, 16.2]	$+1.0$ [-0.3, 3.0]
SC@8^{R_O} − SC@8	$+12.9$ [8.6, 17.7]	$+1.0$ [-0.3, 3.0]
E4: SC-vote inversions / records flagged by $n_{\text{dist}} \geq 4$	0 / 0 of 112	1 / 1 of 259, on different records
accuracy Δ, abstention policy / SC@2 with R_U	eight-call at 4,096 vs. SC@2 at 16,384 -31.7 [-39.2, -24.4] / -33.9 [-41.5, -26.6]	
tokens per record	30.0k / 6.1k vs. 20.4k / 0.26k	

Table 9: Registered same-cap comparisons with pooled SC samples, audited evaluation, 100 questions per population. SC@16 averages the three pairs of seeds; SC@24 uses all 24 SC samples of a question; the portfolios are averaged over the three seeds. Δ in points with question-clustered 95% CI; the primary endpoint, full portfolio − SC@16, uses Holm correction over the three populations, and its null hypothesis is also rejected in all three under the automatic evaluation, at $+5.3$, $+7.0$ and $+7.7$. Completion tokens per question. No SC answer: share of questions without an SC answer among k pooled samples, observed, with $\mathbb{E}_q\, \tau_q^k$ in parentheses. The full portfolio includes backward cloze, which is stored as an aggregate of its four calls and scored by the answer it returned, as stated in Section 5; the eight-call rows do not use it.

	Original	**Prospective 4,096**	**Prospective 16,384**
full − SC@16, primary, Holm	$+5.7$ [3.0, 8.7] reject	$+9.3$ [5.0, 14.3] reject	$+8.3$ [3.7, 13.7] reject
eight-call − SC@16	$+5.0$ [1.6, 8.7]	$+3.4$ [-0.4, 7.6]	$+9.3$ [4.0, 15.0]
eight-call − SC@24	$+2.0$ [-1.5, 5.8]	$+0.7$ [-3.2, 5.1]	$+9.0$ [3.7, 15.0]
full − SC@24	$+2.7$ [-0.3, 5.7]	$+6.7$ [2.0, 11.7]	$+8.0$ [3.0, 13.3]
SC@16 / SC@24 accuracy	.720 / .750	.433 / .460	.887 / .890
SC@16 / SC@24 tokens	54.4k / 81.6k	60.8k / 91.2k	163.4k / 245.1k
No SC answer, $k = 8$	.32 (.33)	.61 (.63)	.13 (.14)
No SC answer, $k = 16$	.28 (.30)	.56 (.59)	.12 (.12)
No SC answer, $k = 24$	.25 (.28)	.54 (.57)	.11 (.12)

Table 10: Registered same-cap SC controls: prospective cohort, 4,096 tokens, 100 questions $\times$ 3 seeds, automatic evaluation. Each arm X regenerates the eight SC samples with all else fixed, giving SC@8^X; eight-call^X uses five of them; the greedy chain and the repairs are those of the cohort, in which eight-call is 0.466, SC@8 0.373, SC calls finished 20%, records without an SC answer 188 and SC completion tokens per record 30.4k. Δ in points with question-clustered 95% CI; the registered endpoints in the first row use Holm correction over the two arms, with the decision on the equality null and the sign of a rejection in parentheses; the registered confirmation requires a rejection with a positive sign.

	Temperature 0.3	**Budget instruction**
eight-call $-$ SC@8^X, registered, Holm	$+8.7$ [4.7, 13.2] reject $(+)$	-11.3 [-17.4, -5.6] reject $(-)$
same under R_O	$+9.4$ [4.7, 14.4]	-6.2 [-11.3, -1.3]
eight-call^X $-$ SC@8^X	$+8.5$ [4.8, 12.7]	-0.2 [-2.3, 2.1]
same under R_O	$+9.6$ [5.6, 14.0]	$+2.1$ [-0.2, 4.8]
SC@8^X $-$ SC@8	$+0.5$ [-2.7, 3.5]	$+20.5$ [14.0, 27.3]
SC@8^X	.378	.578
SC calls finished / records without an SC answer	22% / 186	39% / 122
SC completion tokens per record	30.1k	28.4k

Table 11: Registered run with the budget instruction of Table 10 appended to every prompt, including SC samples, greedy chain and both repairs: prospective protocol, 4,096 tokens, 100 questions $\times$ 3 seeds, automatic evaluation; SC@8 and the eight-call portfolio come from the same run. Δ in points with question-clustered 95% CI; the first row is the registered endpoint, a single test at level 0.05.

eight-call $-$ SC@8, registered	$+9.7$ [5.1, 14.7] confirmed
SC@8 / eight-call	.580 / .677
eight-call $-$ SC@8 of the budget-instruction arm of Table 10	$+9.8$ [5.0, 15.1]
eight-call $-$ eight-call of the cohort without budget instruction	$+21.1$ [14.2, 28.3]
calls finished: SC / greedy / targeted repair / regeneration	38% / 42% / 53% / 45%
records without an SC answer; repair plurality correct on them	123 / 300; 26.0%
completion tokens per record: eight-call / SC@8	28.1k / 28.4k

GSM8K registration. The 1,319-question GSM8K (Cobbe et al., 2021) portfolio was generated under the same protocol with seed 11 at the same budget and evaluated under a separate registration with the frozen configurations; references are the original GSM8K test answers. Because GSM8K answers often carry units, as in "3 bolts", a registered amendment, made after the analysis had been run on the first 60 records, whose output was seen, and before any output on the other records, reduces a located span that is not a plain number to its last number. The primary population excludes the 150 questions of earlier GSM8K analyses, leaving 1,169. Table 7 lists the registered decisions on the automatic evaluation. The registered audit covers 14,954 scored answers. We read the last 700 characters of all 1,054 answers on which a gold-free cross-check disagrees with the located value, the cross-checks being the last box, the last value after the GSM8K final-answer marker, the number after the last answer anchor and the last number. We read the closing 110 characters of the 643 answers selected by a gold-free screen for an alternative form, such as the parenthetical "(or 2 hours and 40 minutes)", or for several numbers in the closing sentence, and again the closing 400 characters of the 609 of them whose final line is not a single number equal to the located value. Of the remaining 13,257 answers, the closing 400 characters were read for the 217 whose final line does not state exactly one number equal to the located value; for the other 13,040 the final line states exactly one number and it equals the located value. Wider context was read wherever the answer sentence lies earlier, including every read answer that closes with a check or verification block. The audit restored the stated answer in 494 answers, 483 of which are correct; in these a heading such as "Total cost:" or an abbreviation such as "Mr." had been located instead of the value, or the last number of a span such as "5 hours to have 60 snowballs", or a closing summary that restates an input after the answer. Eight stop without a final answer, and no credited answer among those read states another value. With these corrections SC@8 is 0.968, and the audited values of the two registered endpoints are $+0.19$ $[-0.13, 0.58]$ for the eight-call and $+0.26$ $[0.00, 0.60]$ for the full portfolio.

Table 12: Registered runs on the 100 prospective questions × 3 seeds, contexts of 12,288 tokens at the 4,096-token cap and 40,960 tokens at 16,384. Runs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Llama-3.1-8B below, at 4,096 and 16,384 tokens and Qwen3.5-9B with the chat template's thinking disabled at 4,096 tokens. Repair variants: the targeted repair regenerated from the logged Qwen3.5-9B greedy chain of every prospective record (TR), the same prompt with the same number of steps marked at random (RM), and an unmasked rewrite with the matched instruction (UR); for Llama-3.1-8B they are candidates of its own run. Candidate-level contrasts: share of records on which the candidate returns a correct answer. Δ in points with question-clustered 95% CI on the registered automatic evaluation; p: two-sided bootstrap sign probability; decision on the equality null, with the sign of a rejection in parentheses: Holm within the registered families, namely the two Llama-3.1-8B eight-call endpoints and the eight repair-variant endpoints, and a single test at level 0.05 for the thinking-disabled endpoint. Audited: after reading every Qwen3.5-9B candidate scored incorrect or unparsed and every Llama-3.1-8B candidate flagged by a gold-free screen, with at least the closing 1,500 characters of every adjudication as "no final answer" re-read; R_O: recovery rule R_O applied to every unfinished candidate.

Run	**Endpoint**	Δ [95% CI]	p	Decision	Audited	R_O
Llama-3.1-8B, 4,096	eight-call − SC@8	-4.2 [-7.8, -0.7]	0.018	reject (−)	-4.7	-4.2
Llama-3.1-8B, 16,384	eight-call − SC@8	-4.2 [-8.3, 0.1]	0.053	retain	-4.2	-4.2
Qwen3.5-9B, thinking off, 4,096	eight-call − SC@8	$+1.2$ [-0.5, 3.8]	0.27	retain	$+1.2$	$+1.0$
Qwen3.5-9B, 4,096	TR − RM	0.0 [-6.0, 6.0]	0.99	retain	$+0.3$	$+3.0$
Qwen3.5-9B, 16,384	TR − RM	-0.3 [-4.7, 4.0]	0.89	retain	0.0	$+0.3$
Qwen3.5-9B, 4,096	TR − UR	-0.7 [-7.0, 5.7]	0.84	retain	-0.7	$+3.3$
Qwen3.5-9B, 16,384	TR − UR	$+12.7$ [8.0, 17.7]	< 0.001	reject (+)	$+13.0$	$+10.7$
Llama-3.1-8B, 4,096	TR − RM	-1.0 [-5.0, 3.0]	0.64	retain	-0.7	-1.3
Llama-3.1-8B, 16,384	TR − RM	-1.3 [-4.7, 2.0]	0.44	retain	-1.3	-1.3
Llama-3.1-8B, 4,096	TR − UR	0.0 [-5.0, 5.0]	1.00	retain	$+0.3$	0.0
Llama-3.1-8B, 16,384	TR − UR	-0.7 [-6.3, 4.7]	0.82	retain	-0.7	-0.7

Descriptive, automatic evaluation with audited values in parentheses:
Llama-3.1-8B, 4,096 / 16,384; Qwen3.5-9B thinking off, 4,096:
SC calls ending within the cap 92% / 96% / 90%; records without an SC answer 0 / 0 / 7 of 300;
SC completion tokens per record 7.5k / 12.4k / 11.6k; SC@8 .417 (.440) / .412 (.427) / .954 (.966);
eight-call under ACDR .375 (.393) / .370 (.385) / .967 (.978);
simple rule .386 (.407) / .377 (.393) / .965 (.977);
full portfolio .383 (.397) / .363 (.373) / .980 (.990).

Descriptive, Qwen3.5-9B repair variants, 4,096 / 16,384 tokens; automatic, audited in parentheses:
correct: TR .333 / .883 (.340 / .890), RM .333 / .887 (.337 / .890), UR .340 / .757 (.347 / .760);
answer returned: TR .340 / .890, RM .337 / .890, UR .347 / .760 under both evaluations;
eight-call with TR / RM / UR in the targeted slot: .469 / .477 / .461 and .984 / .984 / .979
(.471 / .478 / .462 and .987 / .987 / .982);
paired TR − UR in this slot $+0.8$ [-2.3, 3.8] and $+0.5$ [-0.7, 1.7];
regenerated minus original targeted repair in correct answers $+1.0$ [-3.0, 5.0] and $+7.3$ [2.7, 12.0].