SAME SCORE, DIFFERENT STATE: POST-TRAINING MOVES PRE-ANSWER RECOVERABILITY UNDER KV COMPRESSION

Anonymous authors
Paper under double-blind review

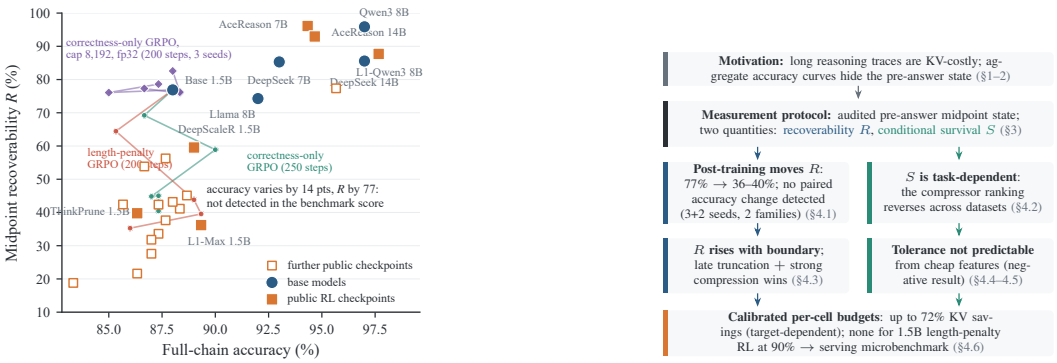
ABSTRACT

Serving a reasoning model means shrinking its key–value (KV) cache while the model is still reasoning. Eviction and a token budget both act on the partial trace before the answer appears, and no benchmark grades that state. We measure it with an answer-marker-audited midpoint protocol: truncate a completed trajectory at 50% of its reasoning tokens, reject prefixes containing answer markers, compress the re-encoded midpoint KV cache, and elicit a fresh boxed answer. The protocol separates midpoint recoverability R, whether a forced answer elicited immediately from the audited pre-answer state is correct, from full-chain accuracy and from conditional eviction survival, with exact KV accounting. Our central finding is that post-training moves R by tens of points while paired full-chain accuracy shows no detectable change, in a recipe-dependent *direction*. Conditional on a correct source trajectory, the Base 1.5B model recovers 77% of midpoint answers on GSM8K. Sixteen public RL checkpoints post-trained from that base, fifteen by five length-shortening recipes, recover 15–60%, all below it, falling monotonically along each recipe's declared dose and tracking realized trace length (Spearman $\rho = 0.96$). The deficit is also present at 8B; on the audited trio it survives a change of elicitation wording, question-only re-solving and, in large part, an answer-leakage audit, and it persists when the model re-reasons from the compressed cache. Paired GRPO arms reproduce it in all twenty runs over three Qwen-1.5B and two Llama-8B seeds. By step 200, R falls to 35–42% at 1.5B under a length penalty and to 45–50% under a correctness-only control, and to 58.5–61.0% at 8B; accuracy is undetectably changed in every run, and the penalty decline is deeper in every 1.5B seed and absent at 8B. Correctness-only arms retrained at an 8,192-token cap with the registered fp32 weights train yet stay at Base level, 17–37 points above their 1,536-cap twins in all three seeds, so this manipulation supports training-time length pressure as the mediator. The direction depends on the recipe: a public RL-only 7B checkpoint recovers 96.1% against its registered base's 85.3% (+10.7 paired, $p = 3.8 \times 10^{-5}$) at indistinguishable accuracy, and a 14B pair repeats it more weakly (+6.8, $p = 6.6 \times 10^{-3}$). Eviction survival is separately model- and task-dependent: compressor rankings reverse across datasets. Under prespecified logistic predictors, no feature family predicts per-trajectory tolerance better than length does. A grouped Clopper–Pearson calibration nevertheless meets per-cell conditional-survival targets on held-out splits (up to 72% KV savings at the 75% retention target, 48% at 90%), while uniform ratios and cross-task transfer both miss it. Two checkpoints with no detectable benchmark-score difference can thus sit 40 points apart in R: same score, different state, in either direction.

1 INTRODUCTION

Reasoning language models trade inference cost for accuracy by generating long intermediate traces (DeepSeek-AI, 2025; Kimi Team, 2025). Every token contributes keys and values at every layer, so long traces constrain batch size and throughput even when weights fit in memory. Whoever serves such a model must decide, per checkpoint and workload, whether to shrink this cache and by how

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.



(a) (b)

Figure 1: (a) Accuracy and midpoint recoverability come apart. On GSM8K, all twenty-six grids, accuracy varies by 14 points while midpoint recoverability R (on correct sources) spans 77. Every length-targeted RL post-trained checkpoint (orange squares) sits below its own base (navy circles), the RL-only AceReason checkpoints *above* their own at both 7B and 14B (Appendix G), and the two GRPO arms move down this axis with no detectable paired accuracy change, while the fp32 cap-8,192 arm (purple diamonds) stays at Base level. Hollow squares are fourteen further public checkpoints (Appendix H); the collapsed α=0.4 run, 75.7% accuracy, is off the plot. (b) The route from one protocol to R, S and calibrated per-cell budgets (gray section pointers).

much. Two lines of work supply the options: post-hoc KV-cache compression (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024; Oren et al., 2024) and training models to reason in fewer tokens (Aggarwal & Welleck, 2025; Hou et al., 2025; Luo et al., 2025a), both graded on one currency, aggregate budget–accuracy curves over completed traces.

That currency scores the wrong state. Cache eviction acts *mid-reasoning*, on a partial trace that benchmarks leave ungraded, so scores leave three deployment questions unanswered. Does a partial trajectory already suffice to recover the answer? How much retained KV can be removed without losing it? Is this *trajectory-level compressibility* cheaply predictable for allocation? This paper makes that state measurable, and the measurement matters: checkpoints with no detectable benchmark difference can differ by 40 points on it before any cache is compressed, in either direction, so a policy read off scores deploys the wrong budget on the wrong model.

Completed-chain evaluation leaks the answer into the compression context; we instead measure a fixed pre-answer-marker point, the first half of reasoning tokens, audited to contain no final-answer markers though it may already contain intermediate results (Section 4.1), separating recoverability from conditional survival.

Our contributions, in order of importance, are:

1. We show that post-training moves midpoint recoverability in a recipe-dependent direction, our central claim. All sixteen public RL checkpoints post-trained from the 1.5B base, and every length-targeted checkpoint at 7B–8B, sit 8–66 points below their base (paired $p \leq 0.011$ on GSM8K, $\leq 1.1 \times 10^{-3}$ on MATH-500; observational), R falls monotonically along three declared-dose ladders (Appendix H), and an RL-only 7B checkpoint sits 10.7 points *above* its base (Appendix G). On the audited 1.5B trio, leakage, elicitation-wording and question-only controls do not account for the gap.

2. We reproduce the erosion at training time under the tested GRPO recipes, with paired arms over three Qwen and two cross-family Llama-8B seeds: R falls from 76.9% to 35.3–42.3% under a length penalty and to 45.0–49.6% under a correctness-only control, with no detectable paired accuracy change (Figure 1) and no detected penalty contrast at 8B. Cap-8192 arms at the registered precision train without eroding in all three seeds, which supports length pressure as the driver (Appendix D).

3. We calibrate one compression ratio per (model, task, compressor) cell with a grouped Clopper–Pearson guarantee on conditional survival (Section 4.5). It meets its target in 48 of 48 held-out cells where uniform ratios miss it in 28 of 39, and it is the paper's deployable output, with savings in a serving microbenchmark.

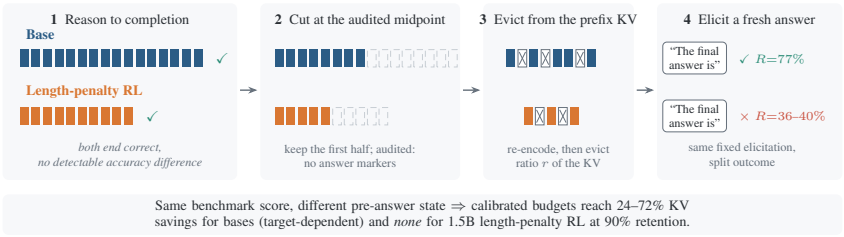


Figure 2: Both models reason to completion and end correct (no detectable paired accuracy difference); each trajectory is truncated at half of its reasoning tokens (audited: no answer markers), the prefix KV re-encoded and compressed at five ratios under three compressors, and a fixed prompt elicits a fresh boxed answer. The Base midpoint still yields the answer; the length-penalty RL midpoint does not (Section 3.1).

4. We register the audited protocol: the truncate-and-force probe of Wang et al. (2026c) made conditional (correct source), audited (no answer marker in the prefix) and inferential (paired McNemar, question-clustered permutation), then extended to KV eviction on the same pre-answer state, where RL-shortened checkpoints degrade faster and do not recover by re-reasoning. It covers thirteen 300-row grids over three base families and three datasets, with a KV-matched oracle-boundary frontier and a prespecified negative result on cheap tolerance predictors as corollaries.

We claim no new compressor, online stopping rule, or universal causal mechanism; our only method-style claim is the calibrated per-cell budget selection. What we claim is actionable measurement: a benchmark score cannot tell which way a checkpoint's pre-answer state has moved, or how much of its cache can be removed safely; the protocol here measures both, the direction of the move tracks training-time length pressure, and calibrated per-cell budgets follow. Two actions follow for practitioners: measure R before adopting a length-targeted recipe, because the benchmark will not show the loss, and keep the completion cap above the base policy's own length distribution, because a lower cap is an undeclared length penalty (Section 5).

2 Related Work

KV-cache compression. Eviction-based compressors discard or merge keys and values to bound inference memory (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024; Oren et al., 2024); later work adapts budgets per query, layer, head or attention pattern (Tang et al., 2024; Cai et al., 2024; Ge et al., 2024; Jiang et al., 2024; Feng et al., 2025) or to reasoning structure (Cai et al., 2025; Wang et al., 2026d; Kai et al., 2026). Benchmarks assess how compression degrades end-to-end reasoning (Liu et al., 2025d;a), with task-dependent compressor behavior on GSM8K already reported for instruction-tuned models (Liu et al., 2025a) and attributed in part to prompt handling (Wang et al., 2026b); replaying a complete trace under a compressed cache can keep the answer while degrading its reasoning (Ai et al., 2026), and Xie (2026) certify eviction-induced failure per cell. A separate line trains *for* compressibility (Gelberg et al., 2026; Du et al., 2025; Zhu et al., 2026); we measure post-training as an uncontrolled *cause* of change in the pre-answer state, and calibrate budgets to a survival target rather than certify error.

Efficient reasoning and its side effects. A growing literature shortens chains of thought: length-penalty RL and pruning (Aggarwal & Welleck, 2025; Hou et al., 2025; Luo et al., 2025a; Arora & Zanette, 2025), token-level compression (Xia et al., 2025; Nayab et al., 2024), latent reasoning (Cheng & Van Durme, 2024; Hao et al., 2024), and inference-time budget control (Muennighoff et al., 2025; Han et al., 2024). Side-effect studies disagree on monitorability (length penalties strip verification and backtracking markers, Little, 2026; no consistent cost, MacDermott et al., 2025); neither measures whether the trace still yields the answer when cut short. Closest to our finding, outcome-reward RL can leave the reasoning prefix causally less important while accuracy rises (Yu et al., 2026), and prefix transferability drops late in RLVR training at stable accuracy (Lee et al., 2026); neither attributes the change to a length term or a completion cap, which our paired arms isolate. The cap is a known pressure: truncated rollouts scored zero shorten traces (Yeo et al., 2025; Yu et al., 2025; He et al., 2025), a pressure He et al. (2025) keep deliberately and defend on end accuracy, Yang et al. (2026) call a constraint rather than a penalty, and Wang et al. (2026a)

attribute to rollout-group composition; we supply the dependent variable those readings lack, and the 8,192-cap arm rules out the length bias of Liu et al. (2025e) as the cause.

Anytime and interrupted reasoning. Truncating a chain of thought and forcing an answer is the probe of Wang et al. (2026c), taken across cut fractions as an area under the curve; interruption studies report reasoning leakage and unpredictable drops (Wu et al., 2025; Koh et al., 2026), the Anytime Index tracks quality against absolute budgets (Zhang et al., 2026), matched absolute horizons show no penalty for prompt-based shortening (Carlon et al., 2026), and training on truncated budgets optimizes anytime accuracy directly (Qi et al., 2025; Xu et al., 2026; Bian et al., 2026). Our R is that probe at a single fraction of each model's own trace, audited for answer markers (a check Koh et al. (2026) note is usually missing), conditioned on a correct source and read with paired inference; the fraction unit, and the reversal at matched absolute budgets (Appendix S), separate our erosion from the absolute-budget results.

3 Measurement Protocol

3.1 Definitions and notation

Let T_i be a completed reasoning trajectory for question q_i with gold answer a_i, and N_i its reasoning-token count (the body before `</think>`). The *midpoint state* $M_i(f)$ is the KV cache obtained by prefilling the question and the first $\lfloor fN_i \rfloor$ reasoning tokens; $f{=}0.5$ unless stated (Figure 2). An elicitation $E(M_i(f))$ appends a fixed instruction and prefills `The final answer is \boxed{`; the prediction is *committed* only if generation closes that box, and *correct* if the boxed content equals a_i. Writing $C_m = \{i : \text{trajectory } T_i \text{ of model } m \text{ is correct}\}$, we define *midpoint recoverability* (the truncate-and-force probe of Wang et al., 2026c at a single fraction, conditioned on a correct source; an anytime accuracy in the sense of Zhang et al., 2026; Qi et al., 2025)

$$R_m(f) \;=\; \frac{1}{|C_m|} \sum_{i \in C_m} \mathbf{1}\big[\, E(M_i(f)) \text{ committed and correct} \,\big]. \tag{1}$$

A sample is *eligible* for model m when its shared ratio-zero elicitation is committed and correct, and conditional quantities are computed over that set: *conditional survival* is

$$S(m,c,r) \;=\; \frac{\big|\{i : \text{eligible}(i,m) \wedge \text{correct}(i,m,c,r)\}\big|}{\big|\{i : \text{eligible}(i,m)\}\big|}. \tag{2}$$

$1 - S$ is the eviction-induced failure rate of Xie (2026), here per (model, compressor, ratio) cell. The per-sample *critical ratio* $r_i^*(c) = \max\{r : \text{correct}(i,m,c,r')\ \forall\, r' \le r\}$ is the largest ratio up to which correctness is preserved contiguously from zero; the maximum observed success $r_i^{\max}(c)$ is retained as a sensitivity quantity for the rare non-monotone recoveries (counted per grid). A sample fails *static preflight* (and is excluded before any grid runs) if its retained prefix contains a boxed-answer marker or the phrase "final answer".

3.2 Deterministic midpoint boundary

For each completed trajectory, we isolate the reasoning body before `</think>`, tokenize it without special tokens, and retain exactly $\lfloor 0.5N \rfloor$ of its N tokens. Five trajectories failed static preflight (counts in the manifest). The question and the retained reasoning are concatenated into one re-encoded context; the original chat roles and the rolling KV state are not preserved (recomputing R and both paired contrasts directly on the rolling state leaves them unchanged; Appendix V). The model receives this context and a fixed instruction to state only the final answer; generation starts after the prefilled string `The final answer is \boxed{` and commits per Section 3.1. The ratio-zero continuation is generated once per sample and shared across compressor records.

3.3 Compression grid and eligibility

We evaluate SnapKV, TOVA, and StreamingLLM (Li et al., 2024; Oren et al., 2024; Xiao et al., 2024) through KVPress 0.5.3 (NVIDIA, 2024) at compression ratios $\{0, 0.25, 0.5, 0.75, 0.9\}$. Every cell uses a new dynamic cache. The recorded cost is the observed retained context-cache length plus the fixed elicitation tokens, not a completion-length proxy.

3.4 REPRODUCIBILITY

Runs use BF16 unless otherwise recorded in the manifest (the seed-7 L1-Max and ThinkPrune pools are FP16; a BF16 L1-Max regeneration shifts R nonsignificantly, 41.2% vs. 36.2%, $p = 0.24$); all Hugging Face model and dataset revisions are pinned; trajectory pools are sampled once at temperature 0.6, top-p 0.95, seed 7, and every elicitation is greedy. Each JSONL row stores schema versions, hashes, revisions, token counts, raw continuations, predictions, correctness and exact KV costs, checked by a validator.

4 RESULTS

Design. One main design, then transfer checks (Appendix Table 3): the protocol, its controls, twenty paired GRPO arms and the completion-cap manipulation all run on GSM8K with the 1.5B DeepSeek-R1-Distill-Qwen base; every other model (7B–14B) and dataset (MATH-500, MMLU-Pro) asks only whether a finding transfers, observationally except for the Llama-8B arms.

4.1 MIDPOINT RECOVERABILITY IS A MODEL PROPERTY WHOSE SIZE DEPENDS ON THE TASK

All thirteen originally registered 300-row grids pass validation (per-grid statistics: Appendix Table 4; GSM8K values, plus the post-hoc AceReason pairs at 7B and 14B, in Figure 1). Confirmatory: the Base/L1-Max/ThinkPrune GSM8K grids, the 7B dataset grids and the prespecified method gate (Section 4.4); the Base MATH-500, Llama-8B, DeepScaleR and Qwen3-8B-pair grids are post-hoc exploratory replications, as is every later grid (MMLU-Pro trio, AceReason pairs, boundary fractions).

The raw eligible rate falls with task difficulty almost entirely because full-chain accuracy falls. $R(0.5)$ conditions *strictly* on correct sources: midpoints from *incorrect* sources also elicit the gold answer (5–29% per grid; Appendix I), and counting them would inflate every rate by 1–26 points.

Three patterns emerge (all paired contrasts: Table 1).

1. **The GSM8K deficit spans every length-related RL checkpoint tested:** 37–41 points below Base for the length-penalty pair (37.6–38.8 measured directly on the rolling KV state, Appendix V; case: Appendix AG) and 17 for DeepScaleR, an outcome-RL comparison trained by iterative context *lengthening* rather than length reduction (Luo et al., 2025b); it persists at 8B, where L1 from Qwen3-8B recovers 87.7% against 95.9% (Aggarwal & Welleck, 2025; Yang et al., 2025). Thirteen further public checkpoints shortened from the same base by five recipes all sit 19–66 points below it ($p \leq 4.3\times10^{-7}$), R falling monotonically along each declared-dose ladder (cap 4k→2k: 53.8→37.6%; α 0→0.4: 41.1→14.5%; target 4,096→1,024: 42.4→21.6%) and tracking realized length across all sixteen (Spearman $\rho = 0.96$), with paired accuracy detectably lower only at $\alpha \geq 0.2$; an SFT-only 7B shortening erodes least (-7.6) and an RL Llama-8B checkpoint by 31 points (Appendix H). The direction depends on the recipe: the RL-only AceReason-7B instead recovers 10.7 points *more* than its own base at equal accuracy, and its 14B counterpart 6.8 points more (Appendix G).
2. **It replicates on the harder MATH-500:** paired gaps of 31.9 points for ThinkPrune and 13.1 for L1-Max, whose recoverability rises even though its full-chain accuracy *drops* from 89.3% to 65.7%.
3. **Even the 7B model is task-dependent:** strong on mathematical tasks but down to 64.1% on MMLU-Pro.

The base-side rate is not Qwen-specific: cross-family Llama-8B lands within three points of Base-Qwen and 35 above the length-penalty pair. Recoverability thus varies with model and task and cannot be read off full-chain accuracy; on the non-mathematical MMLU-Pro the same 1.5B trio shows paired gaps of 42.1 and 45.7 points, wider than on GSM8K (post-hoc, Appendix F).

From association to controlled training-time erosion. We run GRPO *from the registered Base checkpoint itself* in two arms, correctness reward minus $0.2\times\text{tokens}/1000$ and a correctness-only control (α=0), over three training seeds with a fixed data order, evaluating checkpoints through

Table 1: All cross-model paired recoverability contrasts (jointly source-correct; disc. = first-listed / other only; two-sided McNemar exact). The last two GSM8K rows are the post-hoc 7B and 14B RL-only pairs of Appendix G, whose sign is *reversed*.

Contrast	n	recov.	disc.	p
GSM8K				
Base vs. L1-Max	250	190/90	110/10	1.9×10^{-22}
Base vs. ThinkPrune	245	187/96	100/9	1.4×10^{-20}
Base vs. DeepScaleR	247	189/150	60/21	1.7×10^{-5}
Qwen3 vs. L1-Qwen3	288	276/253	30/7	1.9×10^{-4}
DeepSeek vs. AceReason (7B)	271	231/260	10/39	3.8×10^{-5}
DeepSeek vs. AceReason (14B)	280	241/260	13/32	6.6×10^{-3}
MATH-500				
Base vs. L1-Max	175	137/114	35/12	1.1×10^{-3}
Base vs. ThinkPrune	182	140/82	65/7	7.0×10^{-13}

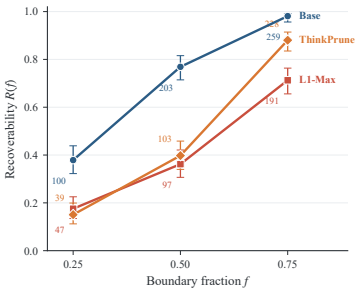
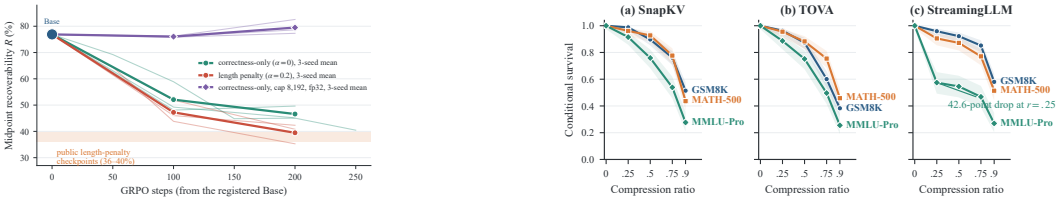


Figure 3: Recoverability $R(f)$ by boundary fraction on correct sources (95% Wilson intervals; Section 4.3).



(a) (b)

Figure 4: (a) Midpoint recoverability versus GRPO steps: three-seed means per arm (bold) with individual seeds (light); every run at the 1,536 cap ends far below Base with no detectable paired accuracy change, and the fp32 cap-8,192 arm (purple) stays at Base level. (b) Conditional survival for DeepSeek-7B across datasets and compressors: StreamingLLM is strongest on GSM8K, tied on MATH-500, weakest on MMLU-Pro.

the registered pipeline. Every run erodes R (paired exact McNemar on jointly correct questions): the seed-7 penalty arm falls monotonically from 76.9% to 35.3% (p down to 3.1×10^{-23}); at step 200 the penalty arm reaches 35.3–42.3% across seeds and the control 45.0–49.6% (all $p \leq 3.3 \times 10^{-12}$), the penalty arm ending below the control in every seed (-2.8 to -9.8 points, pooled -6.8; question-clustered permutation Eq. 3 $p = 1.1 \times 10^{-3}$; Holm-adjusted $p = 0.023$ at step 200 over the four seed-7 checkpoints); full-chain accuracy shows no detectable paired change in any run (85–91%, paired $p>0.22$; Figure 4; equivalence analysis in Appendix B). The erosion crosses families: identical paired arms from the Llama-8B base over two seeds fall from its 74.3% to 58.5–61.0% by step 200 (all $p \leq 3.3 \times 10^{-5}$; accuracy stable at 89.7–94%), with no detected arm contrast at this scale in either seed. Under the tested GRPO recipes, outcome-RL training consistently erodes pre-answer recoverability across seeds and families; the length penalty deepens the decline at 1.5B but not at 8B ($+0.6$ points pooled, $p = 0.87$; Appendix A), and an SFT control on Base's own traces erodes far less (67.5% vs. 45.0–49.6% at step 200; Appendix C). Three correctness-only arms retrained at cap 8,192 in bf16 sit at Base-level R but never trained (1.8% of parameters changed: Adafactor's 10^{-6} step is below half a bf16 ulp, as an engineering-matched control at the registered cap confirms). Rerun with the registered fp32 weights in the same three seeds, the arms train (89% of parameters changed) and do not erode: R is 75.8–76.2% at step 100 and 77.3–82.6% at step 200 against Base's 76.9% (pooled -0.3 and $+3.5$ points, clustered p=0.96, 0.21), $+17.2$ to $+36.9$ points above the same-seed 1,536-cap twins (pooled $+24.5$ and $+32.9$, $p=1.0 \times 10^{-5}$), at unchanged accuracy. Removing the cap removes the erosion in every seed: training-time length pressure, explicit or implicit (truncation zero-rewarded 17–73% of each registered α=0 batch), is supported as the mediator by a manipulation, with the caveat that the cap also removes most of the reward variance the recipe learns from (77% of prompt groups carry zero advantage against 41%; Appendix D). A bf16 Llama-8B cap-8192 arm likewise stays at its base's R (74.2% against 74.3%; $+15.7$ points above its 1,280-cap twin, $p=3.4 \times 10^{-6}$) but moves only 4.6% of its parameters, so at 8B the manipulation is consistent with the account without testing it (Appendix D). Pooled significance uses a question-clustered sign-flip test over 10^5 sign assignments (the seeds re-evaluate the same questions; Eq. 3, Appendix A). The test is conditional on the realized checkpoints; inference over training randomness rests on three (1.5B) and two (8B) seeds.

Cross-sectionally, the same dissociation holds: within the shared-support length stratum (309–830 tokens) the deficit persists (Base 60.0% versus 35.5%/38.6%; Appendix J). DeepScaleR adds a whole-model counterexample: it applies no length pressure, its median trace length (1,106 tokens) is within 10% of Base's grid median (1,230), yet its recoverability is 17 points lower. The deficit thus persists under stratification, matched pairing (17:6 and 17:4 discordants, $p \leq 0.035$), and a length-matched whole-model comparison, and extends beyond length-targeted training.

A second concern is semantic leakage: the retained prefix may literally contain the computed answer. On GSM8K the gap survives it: on jointly gold-absent, jointly correct questions, paired McNemar still favors Base ($p \leq 1.4 \times 10^{-6}$; frozen specification, Appendix K). On GSM8K the gold string appears in 52%/51% of Base/Llama-8B correct sources but only 21%/16% for L1-Max/ThinkPrune, so those midpoints less often contain the completed computation, and their wrong answers are mostly numbers absent from the whole trace (Appendix L). On MATH-500 the ThinkPrune gap persists nominally on the gold-absent subset; L1-Max's is carried largely by gold presence (Appendix K). Answer presence inflates every rate and leaves the GSM8K gap in place; an identical-rule extension finds the DeepScaleR gap also survives gold-absence ($p = 0.024$, but 0.058 under a separator-tolerant rule; the Qwen3-pair contrast is unresolved there). Wording leaves it in place as well: three equivalent prompts move every rate by at most 3 points with an invariant ordering (Appendix O); a free-elicitation control shows R measures *immediate forced* extraction rather than information absence (given budget to re-reason the deficit persists for L1-Max but reverses for ThinkPrune; Appendix P), and once the cache is compressed, the RL checkpoints stay behind after re-reasoning (their free continuation degrades faster than Base's; Appendix Q), and immediate extraction is the regime in which a compressed midpoint cache pays off without regenerating the trace; under a deployed think-close early exit the RL checkpoints reproduce their forced rates ($p \geq 0.13$) while Base rarely commits, so the forced prefill is the stronger deployable operator (Appendix R). Re-solving from the question alone leaves it in place too: with the same elicitation and an empty prefix all three models recover only 4–6%, while the own prefix adds 28–72 paired points and random or other-question prefixes add nothing (Appendix U).

4.2 Conditional survival decreases with compression; the compressor ranking is dataset-dependent

Appendix Tables 5 and 6 and Figure 4b report conditional survival. Each curve conditions on that model's own eligible set; the Base-versus-RL gap persists on the identical jointly eligible subset (Appendix AE) and under engaged decode-time eviction (Appendix AH). On GSM8K, StreamingLLM has the largest mean critical ratio in every grid (1.5B: 0.59–0.76; 7B–8B: 0.77–0.79). On MATH-500 the three are effectively tied at 75% (77.7%, 75.4%, 77.1%; Appendix E). On MMLU-Pro the ranking reverses: SnapKV retains 53.9% at 75% against StreamingLLM's 46.8%, and StreamingLLM already loses 42.6% of eligible answers at 25% compression, its mean critical ratio 0.411, the lowest of any 7B cell. MMLU-Pro is ten-way multiple choice (chance 10%; incorrect-source rows hit 9–13% at r=0.9; question-only recovers 28.2% vs. 63.5% with the own prefix; Appendices I, U): the 90% cells (25–28%) sit near that floor, but the 34-point SnapKV–StreamingLLM split at 25% far exceeds guessing. Compressor conclusions are therefore task-local, replicating the GSM8K task dependence reported for instruction-tuned models (Liu et al., 2025a) under a conditional metric; since the compressors also differ in how much of the prompt they keep (Wang et al., 2026b), we read the reversal as task locality rather than a ranking. One unexplained outlier (Llama-8B SnapKV at 90%: 22.2% vs. 52.2% for TOVA) and rare non-monotonicity (1–15 per cell) are reported without interpretation.

4.3 Boundary sensitivity: recoverability rises steeply with boundary fraction

For Base 1.5B the full grid was repeated at fractions 0.25 and 0.75: recoverability rises steeply (Figure 3; 106 questions gained against 4 lost from 0.25 to 0.50, $p = 9.2 \times 10^{-27}$; 57 against 1 from 0.50 to 0.75, $p = 4.1 \times 10^{-16}$), so extractable answers concentrate late even for Base.

The cross-model contrast repeats the protocol on L1-Max and ThinkPrune at all three fractions. Both RL curves sit far below Base at every boundary (Figure 3; paired McNemar $p \leq 2.8 \times 10^{-7}$, all six). The prespecified slope contrast ($\Delta R_m = R_m(0.75) - R_m(0.25)$; Appendix W) finds L1-Max indistinguishable from Base ($p = 0.21$) and ThinkPrune steeper (73.9 vs. 61.6 points on the jointly source-correct subset, $p = 0.0035$, surviving correction). Both are lower everywhere in the trace; ThinkPrune additionally concentrates its recoverable content later, and at matched fixed

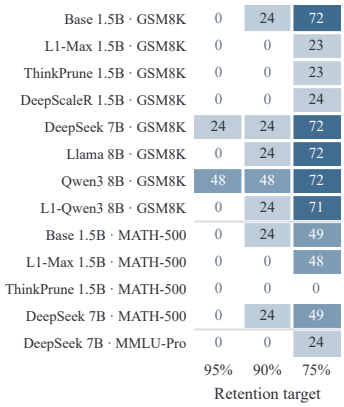


Figure 5: Held-out calibrated KV savings (%), best compressor per cell; "0" = no compression (safe fallback). The blank 1.5B RL block is the deficit as serving policy.

token budgets the ordering reverses, as predicted (Appendix S). The RL-only 7B public pair runs the other way at every boundary: AceReason-7B sits *above* its base by $+33.6$ points at $f{=}0.25$ and $+10.7$ at $f{=}0.5$ (both models at a ceiling by 0.75), so its curve is flatter (Appendix X).

Each (fraction, compressor, ratio) policy is a point in the KV–accuracy plane with exact KV accounting; late truncation with strong compression dominates the uncompressed midpoint for every model (oracle-boundary measurement, Bonferroni over 135 pairs; Appendix Y).

4.4 Per-trajectory tolerance is wide but not cheaply predictable (prespecified negative result)

Spearman correlations between midpoint length and critical ratio are small in 33 of 39 cells (permutation $p > 0.05$, no positive association surviving Bonferroni), yet per-trajectory distributions are wide and sometimes bimodal (Appendix AD): the variance is real. Whether it converts is settled by a prespecified *method gate*, a paired-bootstrap Brier screen against a length-only baseline, then fixed-budget knapsack allocation over observed pre-answer KV, per (model, compressor, nonzero-ratio) cell (Appendix N, Table 24). Over the four GSM8K grids the surface-feature model is better in 0 of 48 cells, worse in 20, and passes 0 of 36 allocation cells; a prespecified internal-feature follow-up (prompt log-probabilities, selected on development data) fails the Brier screen and all but one allocation cell on held-out MATH-500, and a process-signal family fails development screening. The null concerns trajectory-level tolerance from training-free features, a different question from learned token importance. Cheap per-query prediction of tolerance is therefore rejected in every tested setting, with one qualification: post hoc, random forests over hidden states improve Brier in 20–22 of 48 cells yet pass allocation in only 2–3 unstable cells, real signal, unconverted.

4.5 Calibrated compression budgets: a corollary

Per-trajectory prediction fails, but the measurements support grouped budget selection per (model, task, compressor) cell, a corollary rather than adaptation. Each grid's eligible set splits 50/50 into calibration and held-out halves (seed 7); ratios are tested upward, stopping at the first whose one-sided Clopper–Pearson lower bound misses the target $1{-}\alpha$ (fixed sequence, miscoverage $\leq \delta{=}0.05$); if none passes, no compression; the budget is deployed once on the held-out half.

Across 13 grids $\times$ 3 compressors $\times$ 3 targets the rule selects nonzero compression in 48 cells, *all of which meet their retention target on the held-out half*, and falls back to no compression in the other 69; the split is post-hoc, so the guarantee derives from the calibration bound rather than fresh confirmatory data (Appendix AA).

The budgets are sharply model- and task-dependent (Figure 5): at 90% retention the 7B–8B mathematical grids calibrate to 24–48% savings while every 1.5B RL grid selects *no* compression, the deficit as serving policy; at 75% savings reach 72%. The target is the operator's choice: 90% retention suits answer-critical serving, and 75% suits budget-first settings such as bulk scoring or speculative drafting, where a quarter of the recoverable midpoint answers is traded for cache; each target is reported so the trade is explicit. Both natural alternatives fail: uniform $r{=}0.5$ violates the 90% target on 28 of 39 held-out cells, and GSM8K-calibrated budgets miss it on MMLU-Pro and MATH-500;

few-shot shrinkage trades coverage for selections (Appendix AC). Over twelve matched Llama-8B batches the calibrated policy decodes 23.6% faster (CI 21.9–25.2) with 8.1% lower peak memory (7.3–8.9), winning every batch; accuracy is not compared at this sample size (Appendix AB).

5 Discussion

What the score hides, and why it moves. A benchmark score does not fix the pre-answer state. Across the GSM8K grids full-chain accuracy varies by 14 points while midpoint recoverability spans 77 (Figure 1), and the compression certifiable at 90% retention runs from none to 48% (Section 4.5). Post-training moves midpoint recoverability either way: all eighteen length-shortened public checkpoints, in dose order, and all twenty capped arms sit below their base, the AceReason checkpoints above theirs. What moves is the tail behind the answer, the answer-stable tail of Chang et al. (2026): the length-penalty checkpoints and the capped arms cut it from a median 536 tokens to 134–207 and reach the answer slightly earlier in absolute tokens (316–424 against Base's 467), so its first occurrence moves from 0.45 to 0.66–0.72 of a much shorter trace, as Wei et al. (2026b) also report under RL; AceReason-7B keeps a 1,298-token tail and reaches the answer at 0.32 (Appendix M). The midpoint thus lands before the answer in the shortened checkpoints and after it in Base, which is the erosion, and which reverses at matched absolute budgets, where the shortened models are further through their traces (Appendix S), as Carlon et al. (2026) also find for prompt-based shortening. Both are protocol-conditioned quantities, but the *direction* tracks training-time length pressure: the 1,536-token cap zero-rewarded 17–73% of every registered $\alpha{=}0$ batch, and removing it at the registered precision removes the erosion (Appendix D); the representational mechanism stays open. Three consequences follow. (i) Training: set the completion cap above the base policy's own length distribution (45% of the base pool exceeded 1,536 tokens; a lower cap is an undeclared length penalty) and measure R before adopting a length-targeted recipe. (ii) Serving: RL-shortened checkpoints keep their midpoint deficit after re-reasoning once the cache is compressed (13 vs. 2 of 50 questions; Appendix Q); compressor rankings reverse across tasks, so a budget calibrated on one task is kept to that task (Section 4.2); the per-cell calibration is the deployable corollary. (iii) Prediction: the prespecified logistic gate rejects an adaptive method (Table 24; Appendix N).

6 Limitations

This study has limits of scope and design. $R(f)$ and the frontier are oracle-boundary measurements; only the grouped calibration is deployable (Appendix T). The midpoint truncation is a measurement device, and Appendix S bridges to fixed token budgets. The study spans 1.5B–14B (Qwen/Qwen3/Llama; confirmatory grids 1.5B–8B). MMLU-Pro is 7B plus a post-hoc 1.5B trio (Appendix F), and wording controls cover the 1.5B trio. The public comparison is observational: one pool per checkpoint, recipes that differ in data and steps, and dose ladders from one Qwen base. The accuracy equivalence analysis is post-hoc (Appendix B). The arms cover three seeds at 1.5B and two at 8B, one reward form, data order and step budget, and one SFT seed. The fp32 cap-8192 test of the mediator covers three seeds of one base (an 8B bf16 replicate barely trains), and removing the cap also removes most of the learning signal (Appendix D). The recipe dependence rests on two reversed public pairs (AceReason at 7B and 14B, attenuated at 14B). R is conditional on the forced elicitation: free re-reasoning closes (L1-Max) or reverses (ThinkPrune) the gap without compression but not under it (Appendices P, Q). The cache is re-encoded rather than rolling (R replicates on the rolling state; Appendix V). Long-horizon eviction has an exploratory 8–16k pilot only (Appendices AF, AH). Trajectory pools have few seeds, and budgets assume within-benchmark exchangeability. Serving evidence comes from one Llama-8B/L40S/64-token setup.

Ethics Statement

This work uses only public benchmarks and publicly released models; no human subjects, personal data, or new data collection. Experiments were run on single-GPU setups using an NVIDIA L40S, an NVIDIA L40 (rented; the bf16 cap-8192 seed-7/17 arms, the two engineering-matched control arms and the 14B public pair), an NVIDIA A100 80GB (rented; the fp32 cap-8192 arms), or an NVIDIA GeForce RTX 5060 Ti.

REPRODUCIBILITY STATEMENT

All paper-level summaries are generated from the validated grids by deterministic analysis scripts; the schema-v2 manifest records revisions and SHA-256 hashes of every evaluation grid, trajectory pool, and report it registers, and a public invalidation ledger documents every withdrawn artifact. For the longitudinal GRPO arms the manifest records the training script, configuration, and per-checkpoint evaluations; saved checkpoint weights were evaluated and then deleted for disk space. The original Qwen-1.5B seed-7 length-penalty run predates checkpoint hashing and its step-250 checkpoint was lost before saving; the Qwen seed-7 correctness-only run, the seed-17/42 paired arms, all Llama-8B runs, and all six bf16 cap-8192, four control and six fp32 cap-8192 checkpoints (Appendix D) record checkpoint and trainer-state hashes plus full training logs before weight deletion. `src/train_length_rl.py` re-runs any recipe from the pinned base revision, though GPU nondeterminism precludes bit-identical checkpoints; the 21-line vLLM long-cap shim that the cap-8192 arms require is registered in the manifest alongside them. Appendix Table 2 maps each headline claim to its supporting evidence and artifact, and Table 3 the experimental matrix: the full design runs once, on GSM8K with the 1.5B Qwen base, and every other model and dataset is a transfer check.

LLM USAGE STATEMENT

Large language models were used as assistants throughout: writing analysis and training code, editing prose, and, in response to LLM-assisted internal review, suggesting experimental-design additions (control conditions, the longitudinal GRPO replication, multi-seed pools) and statistical revisions (paired tests, multiplicity correction). All reported statistics were produced by deterministic scripts from registered evaluation artifacts, and the authors made all research decisions, verified every suggestion independently, and take responsibility for all content.

REFERENCES

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

Mengting Ai, Jingrui He, and Yue Guo. Does accuracy equal evidence? reasoning faithfulness under kv cache compression. *arXiv preprint arXiv:2608.01631*, 2026.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tingcheng Bian et al. TRiMS: Real-time tracking of minimal sufficient length for efficient reasoning via RL, 2026.

Zefan Cai et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Zefan Cai et al. R-kv: Redundancy-aware kv cache compression for reasoning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Francesca Carlon, Vincent Ginis, and Andres Algaba. Shorter reasoning, earlier answers? an evaluation of reasoning interfaces, 2026.

Kaiyan Chang, Chenwei Zhu, Yingfeng Luo, Yifu Huo, Chenglong Wang, Xiaoqian Liu, Qiaozhi He, Tong Xiao, Zhengtao Yu, and Jingbo Zhu. APR: Penalizing structural redundancy in large reasoning models via anchor-based process rewards. *arXiv preprint arXiv:2602.00760*, February 2026.

Yi-Chang Chen, Feng-Ting Liao, Da-shan Shiu, and Hung-yi Lee. Rethinking dense sequential chains: Reasoning language models can extract answers from sparse, order-shuffling chain-of-thoughts. *arXiv preprint arXiv:2605.07307*, 2026.

Zigeng Chen et al. VeriThinker: Learning to verify makes reasoning model efficient. *arXiv preprint arXiv:2505.17941*, 2025. Checkpoint Zigeng/R1-VeriThinker-7B.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wenjie Du, Li Jiang, Keda Tao, Xue Liu, and Huan Wang. Which heads matter for reasoning? rl-guided kv cache compression. *arXiv preprint arXiv:2510.08525*, 2025.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2407.11550.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. In *International Conference on Learning Representations (ICLR)*, 2024.

Yoav Gelberg, Yam Eitan, Michael Bronstein, Yarin Gal, and Haggai Maron. Training transformers for kv cache compressibility. *arXiv preprint arXiv:2605.05971*, 2026.

Tingxu Han et al. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.

Shibo Hao et al. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, May 2025.

Bairu Hou et al. Thinkprune: Pruning long chain-of-thought of reasoning llms. *arXiv preprint arXiv:2504.01296*, 2025.

Huiqiang Jiang et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jushi Kai, Zhuiri Xiao, Alexandra Birch, et al. Information-aware kv cache compression for long reasoning. *arXiv preprint arXiv:2606.26875*, 2026.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Seunghee Koh, Sungjae Choi, Minchan Kwon, Sunghyun Baek, and Junmo Kim. </think> doesn't stop reasoning: Analysis of spurious cot termination. *arXiv preprint arXiv:2609.03633*, September 2026.

Hyunseok Lee, Soheil Abbasloo, Jihoon Tack, and Jinwoo Shin. Beyond correctness: Learning robust reasoning via transfer. *arXiv preprint arXiv:2602.08489*, February 2026.

Yuhong Li et al. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bryce Little. Length penalties make chain-of-thought less monitorable. *arXiv preprint arXiv:2607.09786*, 2026.

Minghui Liu, Aadi Palnitkar, Tahseen Rabbani, Hyunwoo Jae, Kyle Rui Sang, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang. Hold onto that thought: Assessing kv cache compression on reasoning, 2025a.

Shih-Yang Liu et al. DLER: Doing length penalty right – incentivizing more intelligence per token via reinforcement learning. *arXiv preprint arXiv:2510.15110*, 2025b. Checkpoint nvidia/DLER-R1-1.5B-Research.

Wei Liu, Ruochen Wang, Zhiyuan Yin, et al. Learn to reason efficiently with adaptive length-based reward shaping. *arXiv preprint arXiv:2505.15612*, 2025c. Checkpoints hkust-nlp/Laser-D-L1024-1.5B, Laser-D-L2048-1.5B, Laser-D-L4096-1.5B.

Xiang Liu, Zhenheng Tang, Hong Chen, Peijie Dong, Zeyu Li, Xiuze Zhou, Bo Li, Xuming Hu, and Xiaowen Chu. Semantic integrity matters: Benchmarking and preserving high-density reasoning in kv cache compression, 2025d.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, March 2025e.

Haotian Luo et al. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025a.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://github.com/agentica-project/deepscaler`, 2025b.

Matt MacDermott, Qiyao Wei, Rada Djoneva, and Francis Rhys Ward. Reasoning under pressure: How do training incentives influence chain-of-thought monitorability? *arXiv preprint arXiv:2512.00218*, 2025.

Niklas Muennighoff et al. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Sania Nayab et al. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024.

NVIDIA. Kvpress: Efficient llm key-value cache compression. `https://github.com/NVIDIA/kvpress`, 2024. Version 0.5.3.

Matanel Oren, Michael Hassid, Yossi Adi, and Roy Schwartz. Transformers are multi-state rnns. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Optimizing anytime reasoning via budget relative policy optimization. In *Advances in Neural Information Processing Systems*, 2025.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. In *International Conference on Machine Learning (ICML)*, 2024.

Chen Wang, Hexuan Deng, Yining Zhang, Yuchen Zhang, Jionghao Bai, Zhaochun Li, Ge Lan, and Yue Wang. Implicit compression regularization: Concise reasoning via internal shorter distributions in RL post-training. *arXiv preprint arXiv:2605.07316*, May 2026a.

Heng Wang, Jielin Qiu, Wenting Zhao, Cheng Qian, Liangwei Yang, Jiawei Han, Heng Ji, Silvio Savarese, Shelby Heinecke, and Huan Wang. Random attention: Rethinking KV cache eviction for efficient reasoning. *arXiv preprint arXiv:2609.03430*, September 2026b.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. In *International Conference on Learning Representations (ICLR)*, 2026c. arXiv:2510.01367.

Zihan Wang, Cheng Tang, Lei Gong, et al. Crystal-kv: Efficient kv cache management for chain-of-thought llms via answer-first principle. *arXiv preprint arXiv:2601.16986*, 2026d.

Chengwei Wei, Jung-jae Kim, Longyin Zhang, Shengkai Chen, and Nancy F. Chen. InfoDensity: Rewarding information-dense traces for efficient reasoning. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026a. Checkpoint amao0o0/InfoDensity-DeepSeek-R1-Distill-Llama-8B.

Zihao Wei, Wenjie Shi, Liang Pang, Jingcheng Deng, Shicheng Xu, Shasha Guo, Zenghao Duan, Jiahao Liu, Jingang Wang, Huawei Shen, and Xueqi Cheng. Dynamic rollout editing for reducing overthinking in RL-trained reasoning models. *arXiv preprint arXiv:2606.17890*, June 2026b.

Tsung-Han Wu, Mihran Miroyan, David M. Chan, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. Are large reasoning models interruptible?, 2025.

Heming Xia et al. Tokenskip: Controllable chain-of-thought compression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Peng Xie. Error certificates for KV-cache eviction via randomized design. *arXiv preprint arXiv:2607.21475*, July 2026. Defines eviction-induced failure, the per-cell estimator our S adopts.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. Scalable chain of thoughts via elastic reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.05315. Checkpoints Salesforce/E1-Math-1.5B, E1-Math-7B, E1-Code-14B, E1-AceReason-14B.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025.

Bangji Yang, Hongbo Ma, Jiajun Fan, and Ge Liu. Batched contextual reinforcement: A task-scaling law for efficient reasoning. *arXiv preprint arXiv:2604.02322*, 2026.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.03373.

Qinan Yu, Alexa Tartaglini, Peter Hase, Carlos Guestrin, and Christopher Potts. Outcome rewards do not guarantee verifiable or causally important reasoning, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, March 2025.

Xuanming Zhang, Shwan Ashrafi, Aziza Mirsaidova, Amir H. Rezaeian, Miguel Ballesteros, Lydia B. Chilton, Zhou Yu, and Dan Roth. Budget-aware anytime reasoning with LLM-synthesized preference data, 2026.

Zhenyu Zhang et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rui Zhu, Weiheng Bai, Qiushi Wu, Yang Ren, Haixu Tang, and Yuchu Liu. How to compress kv cache in rl post-training? shadow mask distillation for memory-efficient alignment. *arXiv preprint arXiv:2605.06850*, 2026.

A MULTI-SEED GRPO ARMS

Pooled significance across training seeds uses a question-clustered sign-flip test. With $d_{q,s} \in \{-1, 0, +1\}$ marking which arm alone recovers question q under seed s,

$$D_q = \sum_s d_{q,s}, \qquad T = \sum_q D_q, \qquad p = \Pr_{\varepsilon}\Big(\Big|\sum\nolimits_q \varepsilon_q D_q\Big| \geq |T|\Big), \quad \varepsilon_q \overset{\text{iid}}{\sim} \text{Unif}\{-1, +1\}, \tag{3}$$

Table 2: Claim–evidence map: what this paper claims, on what evidence, at what scope.

Claim	Evidence	Scope	Status
Post-training moves midpoint recoverability	public checkpoints (paired, Tables 1, 15, 16; three dose ladders) + paired GRPO arms + decode-time replication (Appendix AH)	erosion: length-penalty/length-related checkpoints and the tested GRPO recipes (controlled: Qwen-1.5B 3 seeds, Llama-8B 2 seeds); improvement: one RL-only recipe at 7B and 14B; observational: 1.5B–14B, GSM8K/MATH-500/MMLU-Pro	replicated (direction recipe-dependent)
Deficit robust to elicitation mechanism	forced, free, compressed free continuation, early-exit, fixed-budget, decode-time (Appendices P, Q, R, S, AH)	1.5B trio; fraction-matched comparisons. Uncompressed free continuation converges (L1-Max) or reverses (ThinkPrune); the deficit holds in every compressed cell	supported under compression; elicitation-conditioned at ratio zero
Length penalty deepens the decline	paired arms, step-200 contrast	Qwen-1.5B: directional, 2/3 significant; Llama-8B: not detected (MDE 6.3 pts)	scale-dependent
Length pressure drives the erosion	truncation statistics; answer position; public pairs and three public dose ladders (Appendix H); fp32 cap-8192 arms, bf16 cap arms and their engineering-matched control (Appendix D)	1.5B–14B, observational; three manipulated seeds	supported: the fp32 cap-8192 arms train and do not erode in all three seeds (bf16 arms untrained)
Compressor ranking transfers poorly across tasks	GSM8K/MATH-500/MMLU-Pro grids	DeepSeek-7B	supported
Cheap per-query prediction of tolerance	three feature families, method gate (Table 24) + post-hoc nonlinear check	tested settings; hidden states carry unconverted signal	rejected
Calibration selects safe budgets	held-out halves, all 13 grids	conditional survival only	supported
Calibrated budget buys serving capacity	batched-decode microbenchmark (Table 38)	Llama-8B, one configuration	supported

Table 3: Experimental matrix. The full design (protocol, alternative-explanation controls, paired GRPO arms, the completion-cap manipulation) runs once, on GSM8K with the 1.5B Qwen base (first row); every other cell asks only whether the finding transfers. Observational cells are 300-row midpoint grids on public checkpoints; trained cells are GRPO arms we train from the base and evaluate through the same pipeline. Dashes mark cells with no training run.

Base family, scale	Dataset	Observational (public checkpoints)	Trained from the base (causal)
Qwen 1.5B (main design)	GSM8K	Base, L1-Max, ThinkPrune, DeepScaleR, plus 13 shortened checkpoints in three dose ladders (ThinkPrune 4k/3k/2k, α 0–0.4, Laser-D 4096/2048/1024) and two single recipes (E1-Math, DLER); controls: elicitation wording, question-only re-solving, leakage audit, deployment-style early exit, free and compressed continuation, rolling-cache replay; boundary fractions 0.25/0.75	20 paired GRPO runs (penalty vs. correctness-only, 3 seeds); SFT-on-own-traces control; completion-cap manipulation (3 bf16 arms, untrained; 3 fp32 arms, seeds 7/17/42); engineering-matched control
Qwen 1.5B	MATH-500	Base, L1-Max, ThinkPrune	—
Qwen 1.5B	MMLU-Pro	Base, L1-Max, ThinkPrune	—
Qwen 7B	GSM8K, MATH-500, MMLU-Pro	Base 7B on three datasets; AceReason-7B pair (GSM8K, boundary fractions); VeriThinker-7B SFT-only pair (GSM8K)	—
Qwen3-8B	GSM8K	L1 from Qwen3-8B vs. its base	—
Llama-8B	GSM8K; AIME/AMC pilot	Base (GSM8K); InfoDensity-8B pair (GSM8K); exploratory long-trace pilot	4 paired GRPO runs (2 seeds, cap 1,280); cap-8192 α=0 arm (bf16, 1 seed, 4.6% of weights moved)
Qwen 14B	GSM8K	AceReason-14B pair	—

with 10^5 sign assignments and a fixed generator seed; question-level flips preserve within-question correlation under arm exchangeability.

Table 7 reports the paired GRPO arms per training seed (identical configuration per arm; data order fixed to the seed-7 order; checkpoints evaluated through the registered pipeline at ratio zero). Every cell declines substantially from Base's 76.9% (paired McNemar $p \leq 2.2\times10^{-5}$ at step 100, with five of six cells at $\leq 1.7 \times 10^{-10}$, and $p \leq 3.3 \times 10^{-12}$ at step 200), while source accuracy stays at 86–91% (arm-minus-base within ±3 points in every run; all paired $p > 0.22$, so no test rejects equality; a post-hoc equivalence analysis with margins set after the fact places 17 of 20 runs within ±5 points and bounds every 90% interval inside ±6.1, Appendix B). The direction of the arm contrast at step 200 is replicated in all three seeds, although the contrast is individually significant in only two; pooling discordants over seeds gives a single estimate of the penalty-specific effect: -6.8 points at step 200 (87:138 discordants of 755) and -4.2 points at step 100 (111:143 of 753). Because the seeds re-

Table 4: The thirteen midpoint-v1 grids. Eligible = uncompressed midpoint elicitation committed and correct; $R(0.5)$ conditions strictly on correct source trajectories (numerator in parentheses).

Model	Dataset	Correct src. /300	Eligible /300	$R(0.5)$
Base 1.5B	GSM8K	264	212	76.9% (203)
L1-Max 1.5B	GSM8K	268	102	36.2% (97)
ThinkPrune 1.5B	GSM8K	259	105	39.8% (103)
DeepScaleR 1.5B	GSM8K	267	161	59.6% (159)
Llama-8B	GSM8K	276	207	74.3% (205)
Qwen3-8B	GSM8K	291	279	95.9% (279)
L1-Qwen3-8B	GSM8K	293	258	87.7% (257)
Base 1.5B	MATH-500	193	160	76.2% (147)
L1-Max 1.5B	MATH-500	197	139	60.4% (119)
ThinkPrune 1.5B	MATH-500	203	94	42.4% (86)
DeepSeek-7B	GSM8K	279	243	85.3% (238)
DeepSeek-7B	MATH-500	213	179	77.0% (164)
DeepSeek-7B	MMLU-Pro	156	141	64.1% (100)

Table 5: GSM8K conditional answer survival (%) on each model's eligible set. Columns give compression ratio.

Model	Compressor	.25	.50	.75	.90
Base	SnapKV	94.3	81.1	67.9	52.4
Base	TOVA	96.7	84.0	62.3	36.8
Base	StreamingLLM	94.3	92.0	85.8	58.0
L1-Max	SnapKV	91.2	80.4	64.7	31.4
L1-Max	TOVA	84.3	69.6	47.1	28.4
L1-Max	StreamingLLM	85.3	78.4	65.7	47.1
ThinkPrune	SnapKV	89.5	78.1	50.5	17.1
ThinkPrune	TOVA	89.5	67.6	49.5	29.5
ThinkPrune	StreamingLLM	87.6	77.1	60.0	36.2
DeepScaleR	SnapKV	84.5	75.2	59.0	39.1
DeepScaleR	TOVA	91.9	73.9	49.1	26.1
DeepScaleR	StreamingLLM	88.8	82.6	72.0	53.4
DeepSeek-7B	SnapKV	98.8	89.7	76.1	51.4
DeepSeek-7B	TOVA	96.3	87.2	60.1	38.3
DeepSeek-7B	StreamingLLM	95.9	92.2	85.2	58.0
Llama-8B	SnapKV	98.1	90.3	75.8	22.2
Llama-8B	TOVA	98.6	92.8	75.8	52.2
Llama-8B	StreamingLLM	95.7	94.7	85.5	58.5
Qwen3-8B	SnapKV	97.8	85.3	67.7	35.1
Qwen3-8B	TOVA	96.8	82.1	55.2	35.5
Qwen3-8B	StreamingLLM	98.6	97.8	90.7	54.1
L1-Qwen3-8B	SnapKV	96.9	86.4	74.0	26.4
L1-Qwen3-8B	TOVA	95.7	88.8	64.3	42.2
L1-Qwen3-8B	StreamingLLM	95.3	92.6	84.5	58.9

evaluate the same fixed question set, summed pairs are not independent units; significance therefore uses the question-clustered sign-flip permutation test of Eq. 3 (add-one estimator): $p = 1.1 \times 10^{-3}$ at step 200 ($T = -51$ over 134 active questions) and $p = 0.053$ at step 100 (machine-readable reports: `paper/seed_arms_results.json`, `paper/power_mde_results.json`).

The cross-family replication trains the identical paired arms from DeepSeek-R1-Distill-Llama-8B over two training seeds (7 and 17; bf16 policy; steps 100/200 as independent runs): against its registered base grid (92.0% accuracy, R 74.3%), the control reaches 66.5–67.0%/58.5–59.3% and the penalty arm 65.1–70.2%/58.5–61.0% at steps 100/200 (paired McNemar vs. base at step 200: all $p \leq 3.3 \times 10^{-5}$; at step 100 individually nonsignificant in three of four runs, $p \in \{0.0097, 0.067, 0.075, 0.31\}$), with source accuracy 89.7–94% (paired $p \geq 0.18$) and mean lengths 1397–1622 tokens (seed 7). No additional penalty effect is detected in any run ($p \geq 0.53$); pooled over seeds the contrast is $+0.8$/$+0.6$ points at steps 100/200 (91:87 of 524 and 70:67 of 525 discordants; question-clustered permutation as in Eq. 3, $p = 0.83$/0.88), with 80% power against arm differences of 7.1/6.3 points under an independence approximation (clustering across seeds makes these optimistic). The 8B data therefore exclude arm effects at the larger end of the 1.5B

Table 6: DeepSeek-7B conditional answer survival (%) across datasets. MMLU-Pro is ten-way multiple choice (chance survival 10%).

Dataset	Compressor	.25	.50	.75	.90
GSM8K	SnapKV	98.8	89.7	76.1	51.4
GSM8K	TOVA	96.3	87.2	60.1	38.3
GSM8K	StreamingLLM	95.9	92.2	85.2	58.0
MATH-500	SnapKV	96.1	92.7	77.7	43.6
MATH-500	TOVA	95.5	88.3	75.4	45.8
MATH-500	StreamingLLM	90.5	87.2	77.1	51.4
MMLU-Pro	SnapKV	91.5	75.9	53.9	27.7
MMLU-Pro	TOVA	88.7	75.2	49.6	25.5
MMLU-Pro	StreamingLLM	57.4	54.6	46.8	27.0

range (2.8–9.8 points) but not the smaller. Reports: `paper/llama_arms_results.json`, `paper/llama_arms_s17_results.json`, `paper/power_mde_results.json`.

Table 7: Ratio-zero recoverability R (%) by training seed and arm; Δ_{200} = penalty minus control at step 200, with its within-seed paired exact McNemar p on jointly correct questions.

Seed	Control R_{100}	Control R_{200}	Penalty R_{100}	Penalty R_{200}	Δ_{200}	p_{200}
7	58.9	45.0	43.8	35.3	-9.8	0.0076
17	48.1	49.6	51.9	40.8	-8.8	0.0275
42	49.2	45.1	46.0	42.3	-2.8	0.48

B EQUIVALENCE CHECK ON FULL-CHAIN ACCURACY (POST-HOC)

The paired accuracy tests in Section 4.1 fail to reject a null; they do not by themselves establish that the arms score the same. This appendix reports the corresponding equivalence analysis. It is *post-hoc*: the margins were chosen after the accuracy results were seen, so the tests characterize the precision of the accuracy comparison rather than confirm a preregistered equivalence claim.

For all twenty GRPO runs (Qwen-1.5B seeds 7/17/42 and Llama-8B seeds 7/17, each $\times$ two arms $\times$ steps 100/200) we take the paired full-chain accuracy difference against the registered base grid on the same questions, form a 90% normal-approximation interval for paired proportions, and run two one-sided tests at α=0.05 (Table 8).

- At a $\pm$5-point margin, **17 of 20** runs are statistically equivalent to their base. The three exceptions are all Qwen-1.5B cells whose intervals are simply wide ($+2.7$ with CI $[-0.5, +5.9]$; $+2.0$ with $[-2.0, +6.0]$; -2.0 with $[-5.4, +1.4]$); in two of the three the *arm* scores higher.
- At a $\pm$3-point margin only 4 of 20 pass. We report this plainly: 300 paired questions cannot resolve a 3-point difference, and no claim at that resolution is made anywhere in the paper.
- The observed differences carry *no systematic direction*: 9 negative, 6 positive, 5 exactly zero, spanning -2.33 to $+2.67$ points (sign test on the fifteen nonzero runs, $p = 0.61$).

The last point is the substantive one. Accuracy differences are small and unsigned, while every one of these same twenty runs erodes R, by tens of points and always downward. The dissociation is therefore not an artifact of an underpowered accuracy test placed beside a well-powered recoverability test: the two quantities differ in effect size and in directional consistency, not only in whether a test rejects. What the data support is that the arms sit within a few points of base on accuracy, not that they are identical; the paper states the accuracy result as *no detectable paired change* throughout, and this appendix bounds how large an undetected change could be.

C SFT-ON-OWN-TRACES CONTROL: THE EROSION IS CONCENTRATED IN OUTCOME-RL

To separate outcome-RL from on-distribution training per se, a control arm applies supervised fine-tuning to the same Base checkpoint on its own correct traces: 4 samples per question at the

Table 8: Post-hoc TOST on paired full-chain accuracy, arm minus base, over the same questions ($n = 300$ pairs per run; 90% normal-approximation intervals, equivalently two one-sided tests at α=0.05). Margins were not prespecified.

Model	Arm / step	Δ (pts)	90% CI	±5	±3
Qwen-1.5B					
seed 7	α=0 / 100	$+2.00$	$[-2.02, +6.02]$	no	no
seed 7	α=0 / 200	-0.67	$[-3.67, +2.34]$	yes	no
seed 7	penalty / 100	$+1.00$	$[-2.42, +4.42]$	yes	no
seed 7	penalty / 200	-2.00	$[-5.37, +1.37]$	no	no
seed 17	α=0 / 100	$+0.00$	$[-3.38, +3.38]$	yes	no
seed 17	α=0 / 200	$+0.00$	$[-3.47, +3.47]$	yes	no
seed 17	penalty / 100	$+0.00$	$[-2.80, +2.80]$	yes	yes
seed 17	penalty / 200	$+2.67$	$[-0.52, +5.85]$	no	no
seed 42	α=0 / 100	$+0.00$	$[-3.80, +3.80]$	yes	no
seed 42	α=0 / 200	$+1.33$	$[-2.22, +4.88]$	yes	no
seed 42	penalty / 100	-0.33	$[-4.09, +3.43]$	yes	no
seed 42	penalty / 200	$+1.00$	$[-2.24, +4.24]$	yes	no
Llama-8B					
seed 7	α=0 / 100	-1.33	$[-3.90, +1.24]$	yes	no
seed 7	α=0 / 200	-2.00	$[-4.32, +0.32]$	yes	no
seed 7	penalty / 100	-0.33	$[-2.72, +2.06]$	yes	yes
seed 7	penalty / 200	$+2.00$	$[-0.68, +4.68]$	yes	no
seed 17	α=0 / 100	$+0.00$	$[-2.45, +2.45]$	yes	yes
seed 17	α=0 / 200	-1.33	$[-3.78, +1.12]$	yes	no
seed 17	penalty / 100	-0.33	$[-2.46, +1.79]$	yes	yes
seed 17	penalty / 200	-2.33	$[-4.84, +0.17]$	yes	no
Pass rate				17/20	4/20

trainer-default temperature (1.0) over the same 2,000-question GSM8K train pool yield 1,487 correct traces (completion cap 1,536), on which completion-only SFT runs 200 steps at 6 prompts per step (lr 10^{-5}; the MLE objective and SFT-conventional learning rate are the deliberate differences from the GRPO arms, which are otherwise matched in base, data pool, completion cap, steps, and prompts per step). Checkpoints at steps 100/200 pass through the registered midpoint pipeline, paired against the registered Base grid (script `src/train_sft_control.py`; report `paper/sft_control_results.json`; specification frozen before any run).

The SFT arm erodes mildly: R falls from 76.9% to 67.6%/67.5% at steps 100/200 (paired 57:34 and 49:26 discordants, $p = 0.021/0.011$), with source accuracy stable at 85.0–86.3%. At the same step count the GRPO arms fall to 45.0–49.6% (correctness-only) and 35.3–42.3% (length penalty). Training on the model's own correct traces thus accounts for roughly 9 points of a 27–42-point drop, a real but minor component, and the bulk of the erosion is specific to the outcome-RL arms rather than to continued training as such (single seed; single reward-free objective).

D Raising the completion cap: bf16 arms that did not train, and the fp32 arm that did

Appendix M shows that every controlled arm in this paper trained under heavy truncation: at `--max-completion-length` 1536, the logged `clipped_ratio` of the three α=0 arms ranges over 0.167–0.728, so between a sixth and three quarters of every on-policy batch was cut off before its boxed answer and earned reward zero. A truncated completion is punished exactly like a wrong one, so the cap is an implicit length penalty even in a correctness-only control. This appendix reports the arms that were meant to remove it: three bf16 replicates whose Base-level R an engineering-matched control shows to be uninformative (under the bf16 master weights the long budget forced, the policy does not train), and a fourth arm at the registered fp32 precision that trains and does not erode.

Protocol. Each arm retrains from the same registered Base 1.5B checkpoint with the identical recipe of the registered α=0 arms, α=0, data seed 7, 6 generations $\times$ 6 prompts, lr 10^{-6}, the same 2,000-question GSM8K train pool, 200 steps, checkpoints at 100 and 200, and one sub-

stantive change, `--max-completion-length` 1536 → 8192. The manipulation is replicated over the same three training seeds as the registered arms (7, 17, 42): only `--seed` differs between them, `--data-seed` stays 7 so all three see the registered data order, and seeds 17 and 42 reuse the batch/vLLM geometry the seed-7 memory probe adopted rather than re-deriving it. All six checkpoints pass through the registered midpoint pipeline unchanged (320 seed-7 GSM8K test trajectories at 12,288 tokens, answer-marker preflight at f=0.5 excluding 0–4 of 320 rows per checkpoint, ratio-zero SnapKV on 300 rows, 64-token forced elicitation, seed 7, bf16), are SHA-256-hashed and are then deleted. Runners `run_cap8192_arm.sh` and its `_s17`/`_s42` replicates, report `paper/cap_arm_results.json`, registered as `cap8192_alpha0_s{7,17,42}_v1` / `cap_arm_v1`. Seeds 7 and 17 ran on one rented L40, seed 42 on the registered L40S. A fourth arm repeats the seed-7 recipe with fp32 master weights (`run_cap8192_fp32_arm.sh`; report `paper/capfp32_arm_results.json`, registered as `cap8192fp32_alpha0_s7_v1` / `capfp32_arm_v1`); its fp32 states beside the 8,704-token logits exceed 48 GB, so it ran on a rented A100 80 GB.

The 1,536 cap was not a free parameter of the registered runs. TRL 0.23 hardcodes `max_num_batched_tokens`=4096 in its colocated vLLM construction (a workaround for the V1 memory profiler), and vLLM 0.7.3's V0 scheduler refuses any `max_model_len` above that value with chunked prefill off; with a 512-token prompt budget, 512+1536=2048 is the largest power-of-two completion budget that fits, and 512+8192=8704 aborts at trainer construction. *No arm in the registered matrix could have been trained at a longer budget on this stack.* A 21-line shim (`logs/cap8192_arm/vllm_longcap_shim.py`, registered) raises that one engine argument to `max_model_len` and changes nothing else; `src/train_length_rl.py` and the environment are byte-identical, and the shim cannot fire at cap 1536 because $4096 \geq 2048$.

Three engineering deviations, shared by all three seeds, were forced by the larger logits tensor and are disclosed in the runner headers and the manifest. (i) `--per-device-batch` 1 instead of 4: gradient accumulation is derived as prompts × generations/per-device-batch, so the optimizer step is unchanged at 36 completions per update at lr 10^{-6}, though accumulating 36 bf16 micro-steps is not numerically identical to accumulating nine. (ii) `--vllm-util` 0.35 with sleep mode instead of 0.22 without: the colocated engine must hold 36 sequences of 8,704 tokens (≈8.4 GiB) in its KV cache and is offloaded between generation phases; this changes memory scheduling, not the objective. (iii) bf16 weights loaded directly, where the registered arm trained in fp32 and was cast to bf16 in place before evaluation, the same evaluated dtype, one fewer step. A 12-step probe ladder selected (i) and (ii) before any of the 200 steps was committed, and it was reused unchanged by the two replicate seeds. The fp32 arm keeps (i) and (ii) and drops (iii).

These three deviations are confounds, and until the control below was run they were uncontrolled within the 1.5B matrix: no arm paired the registered 1,536 cap with this configuration, so the three seeds ruled out training randomness without separating the cap from the configuration it forced. The control arms of the next section pair exactly that: the registered cap under the cap-arm configuration.

Result of the bf16 cap-8192 arms: Base-level R in every seed. Table 9 gives all six bf16 checkpoints (its last block is the fp32 arm, read below). R is 77.6–79.7% at step 100 and 74.7–78.9% at step 200 against Base's 76.9%; paired by question text on jointly source-correct questions, *no* checkpoint differs from Base (gaps -0.8 to $+3.7$ points, exact McNemar $p \geq 0.34$), while each seed's own 1,536-cap α=0 twin sits 21.5–31.4 points below it (all $p \leq 5.2 \times 10^{-9}$); every cross-seed twin gives the same picture ($+18.9$ to $+32.5$ points, all $p \leq 9.7 \times 10^{-7}$).

Because the three seeds re-evaluate the same fixed question set, the pooled contrasts use the question-clustered sign-flip permutation test of Eq. 3 (10^5 draws, seed 7) rather than treating summed discordant pairs as independent. Pooled over the three seeds, the cap arms sit $+26.4$ points above their own twins at step 100 (240:42 discordants over 749 jointly source-correct pairs, net $+198$ over 160 active questions) and $+30.4$ at step 200 (266:41 over 740, net $+225$ over 165); the clustered p is at the 10^5-draw floor in both, as is the naive summed binomial (1.5×10^{-12}, 1.7×10^{-12}). Against Base the pooled gap is $+1.8$ and $+0.8$ points (114:101 and 101:95) at clustered $p = 0.57$ and 0.80, null, in the direction of the cap arm, at three times the per-seed resolution.

Full-chain accuracy is 275–281/320 = 85.9–87.8% and is indistinguishable from Base on the 319 shared pool questions in every checkpoint ($p \geq 0.61$), so the "same score" half of the comparison holds here too. Median reasoning length lands at 1013–1114 tokens against Base's 1128, where the

1,536-cap twins had collapsed to 616–739: the long-budget arms neither shorten their chains nor lose recoverability. Read on its own this looked like the cap removing the erosion; the control below shows that these arms had not trained.

Table 9: The cap-8192 correctness-only arm, three bf16 seeds and three fp32 seeds, against the registered Base grid and each seed's own 1,536-cap twin. $R(0.5)$ conditions on correct sources (recoverable = ratio-zero elicitation committed and correct); Δ and p are paired by question text on jointly source-correct questions (two-sided exact McNemar). N = median reasoning tokens on source-correct rows. Accuracy is the 320-row source pool for the cap-arm rows and the 319 shared questions for the Base row; its paired test against Base uses those 319 questions. The pooled rows sum the per-seed pairs and test them with the question-clustered sign-flip permutation of Eq. 3 (10^5 draws, seed 7); "src" is then the pooled number of jointly source-correct pairs *for the twin contrast* (the Base contrast in the same row rests on 737 and 731 pairs), and $p^\dagger$ is at the draw floor.

Checkpoint	step	src	$R(0.5)$	95% CI	acc.	N	vs. Base / vs. own twin
Base 1.5B (registered)	—	264	76.9%	[71.4, 81.6]	87.1%	1128	—
cap-8192 α=0 s7	100	261	79.7%	[74.4, 84.1]	85.9%	1046	$+3.7$, p=0.34 / $+21.5$, 5.2×10^{-9}
cap-8192 α=0 s7	200	259	76.4%	[70.9, 81.2]	85.9%	1042	$+0.4$, p=1.00 / $+30.7$, 1.4×10^{-15}
cap-8192 α=0 s17	100	264	77.7%	[72.3, 82.3]	87.8%	1114	$+0.0$, p=1.00 / $+28.5$, 1.4×10^{-13}
cap-8192 α=0 s17	200	261	78.9%	[73.6, 83.4]	86.6%	1046	$+2.9$, p=0.48 / $+31.4$, 5.3×10^{-15}
cap-8192 α=0 s42	100	259	77.6%	[72.1, 82.3]	86.2%	1027	$+1.6$, p=0.72 / $+29.3$, 4.5×10^{-15}
cap-8192 α=0 s42	200	257	74.7%	[69.1, 79.6]	85.9%	1013	-0.8, p=0.90 / $+29.1$, 1.9×10^{-13}
pooled (3 seeds)	100	749	—	—	—	—	$+1.8$, p=0.57 / $+26.4$, $p^\dagger$
pooled (3 seeds)	200	740	—	—	—	—	$+0.8$, p=0.80 / $+30.4$, $p^\dagger$
cap-8192 α=0 s7, fp32	100	255	76.1%	[70.5, 80.9]	84.4%	1022	-0.4, p=1.00 / $+17.2$, 2.0×10^{-6}
cap-8192 α=0 s7, fp32	200	262	78.6%	[73.3, 83.2]	86.9%	1040	$+2.8$, p=0.47 / $+33.3$, 1.8×10^{-18}
cap-8192 α=0 s17, fp32	100	265	75.8%	[70.4, 80.6]	88.1%	992	-0.8, p=0.91 / $+28.4$, 5.9×10^{-15}
cap-8192 α=0 s17, fp32	200	260	77.3%	[71.8, 82.0]	85.9%	1093	$+1.7$, p=0.70 / $+28.3$, 1.1×10^{-14}
cap-8192 α=0 s42, fp32	100	265	76.2%	[70.7, 81.0]	88.1%	1080	$+0.4$, p=1.00 / $+27.8$, 5.6×10^{-14}
cap-8192 α=0 s42, fp32	200	264	82.6%	[77.5, 86.7]	86.9%	1053	$+6.1$, p=0.077 / $+36.9$, 4.5×10^{-23}
pooled fp32 (3 seeds)	100	746	—	—	—	—	-0.3, p=0.96 / $+24.5$, $p^\dagger$
pooled fp32 (3 seeds)	200	751	—	—	—	—	$+3.5$, p=0.21 / $+32.9$, $p^\dagger$
1,536-cap α=0 s7	100	270	58.9%	[52.9, 64.6]	—	739	—
1,536-cap α=0 s7	200	262	45.0%	[39.1, 51.1]	—	616	—
1,536-cap α=0 s17	100	264	48.1%	[42.1, 54.1]	—	699	—
1,536-cap α=0 s17	200	264	49.6%	[43.6, 55.6]	—	668	—
1,536-cap α=0 s42	100	264	49.2%	[43.3, 55.2]	—	694	—
1,536-cap α=0 s42	200	268	45.1%	[39.3, 51.1]	—	630	—

Truncation statistics. Table 10 reads the TRL logs with the same parser as Table 23. The seed-7 and seed-17 cap-8192 arms never reach their cap in any logged step and the seed-42 arm grazes it once in 40 (largest observed `completions/max_length` 8133, 8143 and 8192 of 8192, i.e. 0–2.5% of steps at cap); their `clipped_ratio` stays in 0.000–0.072, means 0.022–0.027, while every 1,536-cap arm sits at its cap in 95–100% of logged steps at a clipped ratio of 0.106–0.750. Mean completion length runs 1,932–3,174 tokens against 894–1,419: the policy uses the budget it is given, and when it is not punished for doing so it keeps writing chains of Base-like length. This is the manipulation check on the cap: what changed between the two families of arms is how much of each batch the cap destroyed. The fp32 arm (fourth row) reproduces the bf16 arms' exposure, at cap in 0 of 40 logged steps, clipped ratio 0.000–0.078 (mean 0.024), mean length 2,117–3,026, so the manipulation is the same at both precisions. The bf16 control arm at the registered cap (last row) reproduces the registered exposure, at cap in 40 of 40 logged steps, clipped ratio 0.489–0.722, but, unlike the registered arm, its clipped ratio never falls, because its policy never learns to shorten (Table 12).

Answer position. Running Appendix M's Part-A machinery on these arms with the same pinned base tokenizer, none of the six cap-8192 checkpoints back-loads the answer: median f_{first} is 0.436–0.481 against Base's 0.453 (seed 7 0.481/0.447, seed 17 0.441/0.458, seed 42 0.436/0.468 at steps 100/200), and paired on questions with the gold string in both bodies the shift is between -0.011 and $+0.031$ (n=238–246; sign test $p \geq 0.10$, sign-flip $p \geq 0.24$), null in both directions in every seed, where every 1,536-cap arm moved to 0.663–0.720 at $p \leq 6.4 \times 10^{-15}$. The verification tail agrees: 483–578 median tokens written after the answer is first computed, against Base's 536 and 179–190 for the three 1,536-cap α=0 twins. The positional signature of the erosion disappears together with

Table 10: Training-time truncation exposure, from the TRL logs (`src/analyze_cap_arm.py`, block (d)). "at cap" = logged steps whose `completions/max_length` equals the cap. The seed-7 1,536-cap arm ran 250 steps against the cap arms' 200; restricted to the matched first 40 logged steps its clipped ratio is 0.178–0.661.

Arm	cap	logged steps	at cap	clipped ratio (mean)	mean length
cap-8192 α=0 s7	8192	40	0%	0.000–0.061 (0.027)	1960–3108
cap-8192 α=0 s17	8192	40	0%	0.000–0.067 (0.022)	2109–3029
cap-8192 α=0 s42	8192	40	2.5%	0.000–0.072 (0.026)	1932–3174
cap-8192 α=0 s7, fp32	8192	40	0%	0.000–0.078 (0.024)	2117–3026
cap-8192 α=0 s17, fp32	8192	40	2.5%	0.000–0.078 (0.023)	2026–2954
cap-8192 α=0 s42, fp32	8192	40	2.5%	0.000–0.078 (0.023)	1880–3252
1,536-cap α=0 s7	1536	50	98%	0.167–0.661 (0.360)	996–1380
1,536-cap α=0 s17	1536	40	100%	0.200–0.728 (0.393)	1020–1411
1,536-cap α=0 s42	1536	40	100%	0.233–0.672 (0.405)	1062–1395
1,536-cap penalty s17	1536	40	95%	0.106–0.750 (0.351)	894–1419
1,536-cap penalty s42	1536	40	98%	0.133–0.700 (0.357)	971–1417
1,536-cap α=0 s7, bf16 control	1536	40	100%	0.489–0.722 (0.620)	1303–1421

the erosion, in all three seeds, as it must for the bf16 arms, whose policies are at Base (next section). The fp32 arm, whose policy did move, gives the same null: median f_{first} 0.472/0.484 at steps 100/200, paired shift null (sign test p=0.79/0.16, sign-flip p=0.69/0.51, n=233/243), verification tail 488/468 median tokens against Base's 536.

Engineering-matched control at the registered cap. The missing corner of the design is an α=0 arm at the registered cap of 1,536 trained under the cap-arm configuration: bf16 master weights, per-device batch 1 (36 accumulated micro-steps), `--vllm-util` 0.35 with sleep mode, and the byte-identical shim on the code path (inert here, since $4096 \geq 2048$). Everything else, base, data pool and order (`--data-seed` 7), optimizer step, 200-step budget, checkpoints at 100/200, evaluation pipeline, is that of the registered arms, so each control checkpoint shares its cap with the registered twin of its seed and its configuration with the cap-8192 arm of its seed. Runner `run_capctl1536_arms.sh` (seeds 7 and 17), analysis `src/analyze_capctl_arm.py`, report `paper/capctl_arm_results.json`, registered as `capctl1536_alpha0_s{7,17}_v1` / `capctl_arm_v1`; both ran on the rented L40. If the cap were the mediator, this arm should erode like its registered twin (58.9%/45.0% at steps 100/200) and sit far below its cap-8192 sibling. It does neither (Table 11): R is 72.6% and 75.8% at steps 100 and 200 against Base's 76.9% (paired p=0.54 and p=0.72), $+14.8$ and $+30.3$ points above the registered twin (4.8×10^{-5}, 1.8×10^{-16}), and indistinguishable from the cap-8192 arm of the same seed (-6.0, p=0.06; -0.8, p=0.90), at unchanged accuracy (85.6%/87.8%, p=0.51/p=0.85 against Base) and Base-like chain length (median 1004/971 tokens against 1128). Taken at face value this would say that the engineering configuration, not the cap, removes the erosion. The training logs say something more specific.

Table 11: The engineering-matched control: α=0 at the registered 1,536-token cap under the cap-8192 arms' configuration (bf16 master weights, per-device batch 1, `--vllm-util` 0.35 with sleep, shim on the code path). Columns as in Table 9; the last column pairs the control against Base, against its registered 1,536-cap α=0 twin (same cap, registered fp32 configuration) and against the cap-8192 arm of the same seed (same configuration, cap 8,192), by question text on jointly source-correct questions (two-sided exact McNemar).

Checkpoint	step	src	$R(0.5)$	95% CI	acc.	N	vs. Base / vs. 1,536 twin / vs. cap-8192
control α=0 s7	100	259	72.6%	[66.9, 77.7]	85.6%	1004	-2.5, p=0.54 / $+14.8$, 4.8×10^{-5} / -6.0, p=0.06
control α=0 s7	200	264	75.8%	[70.2, 80.5]	87.8%	971	-1.6, p=0.72 / $+30.3$, 1.8×10^{-16} / -0.8, p=0.90

Why the bf16 policy did not move. Table 12 reads the scalars a learning GRPO policy has to move, reward, completion length, clipped ratio, entropy, from every arm's training log, first five against last five logged steps. The registered fp32 arms learn: at the same cap, seed and data order the registered α=0 seed-7 arm raises its reward from 0.46 to 0.78, shortens from 1338 to 1055 tokens, cuts its clipped ratio from 0.59 to 0.21 and its entropy from 0.75 to 0.56. The bf16 control arm moves none of them (reward 0.46 to 0.44, length 1351 to 1368, clipped ratio 0.60 to 0.61, entropy 0.72 to

0.71), and neither do the three cap-8192 arms, whose reward starts near the base policy's accuracy because nothing is truncated and stays there. The cause is arithmetic. `--bf16-weights` loads the trainable parameters in bf16 and the Trainer's Adafactor (`relative_step` off) then applies a nominal per-parameter step equal to the learning rate, 10^{-6}; a bf16 value w has a unit in the last place of $2^{\lfloor \log_2 |w| \rfloor - 7}$, so under round-to-nearest an update of 10^{-6} registers only where $|w| < 4.9 \times 10^{-4}$. That is 1.2% of the model's parameters. The registered arms train fp32 master weights under bf16 autocast, where the same step always registers. A direct comparison of the preserved seed-7 control checkpoint at step 200 with the Base weights (`src/weight_diff_vs_base.py`; `logs/capctl1536_arms/weight_diff_*.json`) confirms the mechanism: 1.78% of the 1.78B parameters differ at all, the mean absolute change is 1.4×10^{-7} against a mean absolute weight of 0.029 (relative change 4.9×10^{-6}), and the changes concentrate where the arithmetic allows them: 80% of the parameters below the threshold changed against 0.9% of the rest. After 200 optimizer steps the control policy is, to within rounding, the Base policy; the cap-8192 checkpoints, trained in the same configuration and with the same flat logs, were deleted after evaluation and could not be diffed, but nothing in their logs distinguishes them from it. The Llama-8B arms (Appendix A) also load bf16 weights and do move, reward $+0.12$ to $+0.23$, length -7 to -9%, clipped ratio 0.62–0.66 $\rightarrow$ 0.45–0.49, and erode; bf16 slows but does not freeze the 8B policies. Why the same precision freezes the 1.5B policy and only slows the 8B ones we have not established (the two families differ in weight magnitudes and gradient norms, $\approx$1 against $\approx$0.3, and no 8B checkpoint was diffed); their erosion rests on logs that do move.

Table 12: Learning dynamics from the TRL training logs (`src/analyze_capctl_arm.py`, block (f)): mean of the first five against the last five logged steps (logging every 5 optimizer steps) for reward, completion length in tokens, clipped ratio and entropy. Master weights are fp32 for the registered 1.5B arms and the fp32 cap-8192 arm and bf16 (`--bf16-weights`) for the control, bf16 cap-8192 and Llama-8B arms; every arm uses Adafactor at lr 10^{-6} with a linear decay.

Arm	weights	cap	reward	mean length	clipped ratio	entropy
1,536-cap α=0 s7 (registered)	fp32	1536	0.46 → 0.78	1338 → 1055	0.59 → 0.21	0.75 → 0.56
1,536-cap α=0 s17 (registered)	fp32	1536	0.39 → 0.73	1384 → 1115	0.65 → 0.29	0.74 → 0.60
1,536-cap α=0 s42 (registered)	fp32	1536	0.46 → 0.73	1363 → 1137	0.59 → 0.30	0.72 → 0.58
1,536-cap penalty s17 (registered)	fp32	1536	0.13 → 0.56	1370 → 1011	0.64 → 0.21	0.75 → 0.56
1,536-cap penalty s42 (registered)	fp32	1536	0.20 → 0.56	1358 → 1042	0.59 → 0.22	0.72 → 0.53
1,536-cap α=0 s7 (control)	bf16	1536	0.46 → 0.44	1351 → 1368	0.60 → 0.61	0.72 → 0.71
cap-8192 α=0 s7, fp32	fp32	8192	0.89 → 0.90	2447 → 2369	0.03 → 0.01	0.79 → 0.76
cap-8192 α=0 s17, fp32	fp32	8192	0.84 → 0.89	2502 → 2580	0.02 → 0.03	0.78 → 0.80
cap-8192 α=0 s42, fp32	fp32	8192	0.87 → 0.89	2516 → 2531	0.03 → 0.02	0.78 → 0.76
cap-8192 α=0 s7	bf16	8192	0.88 → 0.91	2517 → 2385	0.03 → 0.02	0.80 → 0.76
cap-8192 α=0 s17	bf16	8192	0.85 → 0.88	2580 → 2610	0.02 → 0.03	0.78 → 0.79
cap-8192 α=0 s42	bf16	8192	0.88 → 0.90	2481 → 2579	0.03 → 0.03	0.77 → 0.77
Llama-8B α=0 s7	bf16	1280	0.44 → 0.56	1156 → 1075	0.62 → 0.46	0.67 → 0.61
Llama-8B penalty s7	bf16	1280	0.20 → 0.38	1167 → 1076	0.64 → 0.47	0.67 → 0.61
Llama-8B α=0 s17	bf16	1280	0.41 → 0.59	1163 → 1082	0.64 → 0.49	0.66 → 0.61
Llama-8B penalty s17	bf16	1280	0.17 → 0.40	1170 → 1065	0.66 → 0.45	0.67 → 0.60

The cap at the registered precision. The fp32 arms are the same manipulation with the arithmetic fixed: seeds 7, 17 and 42 (data seed 7), cap 8,192, deviations (i) and (ii) retained, fp32 master weights under bf16 autocast as in every registered arm, 200 steps each on a rented A100 80 GB (393–401 minutes; the 48 GB cards run out of memory in the first backward pass). The seed-7 policy moves: diffing the preserved step-200 weights against Base (`logs/cap8192_fp32_arm/weight_diff_*.json`), 89.0% of the 1.78B parameters differ, in all 339 tensors, with mean absolute change 5.6×10^{-6} against a mean absolute weight of 0.029 (relative change 1.9×10^{-4}, forty times the frozen control's), and the changes are uniform across the bf16 threshold (85% below it, 89% above), as fp32 arithmetic predicts. Its curves, however, are as flat as the bf16 arms' (Table 12: reward 0.89 to 0.90, length 2447 to 2369, entropy 0.79 to 0.76), and the training log says why: with almost nothing truncated the reward starts at the base policy's accuracy, and 76% of the prompt groups have all six completions equally rewarded and so carry zero advantage (`frac_reward_zero_std`), against 44% in the registered seed-7 twin, 25% in the control and 6–7% in the penalty arms; the within-batch reward spread is 0.28 against 0.44 and the gradient norm 0.15 against 0.30, with 9 of 40 logged steps at exactly zero. The optimizer registers every update it is given; the cap gives it little to learn. Evaluated through the registered pipeline (Table 9, last block), R is 76.1% (194/255) at step 100 and 78.6% (206/262) at step 200

against Base's 76.9%: paired, -0.4 (33:34, p=1.00) and $+2.8$ (38:31, p=0.47) against Base, $+17.2$ (60:18, $p=2.0\times10^{-6}$) and $+33.3$ (92:9, $p=1.8\times10^{-18}$) above the registered 1,536-cap twin of the same seed, and indistinguishable from the bf16 cap-8192 arm (-3.7, p=0.30; $+2.8$, p=0.41) and the control ($+1.2$, p=0.82; $+3.2$, p=0.36); accuracy 84.4%/86.9% (p=0.21/1.00 against Base), median reasoning length 1022/1040 tokens against 1128, answer position at Base (above). Seeds 17 and 42 repeat every reading: 88.9% and 89.1% of parameters change, 78% and 76% of prompt groups carry zero advantage, and R is 75.8%/77.3% and 76.2%/82.6% at steps 100/200 (-0.8/$+1.7$ and $+0.4$/$+6.1$ against Base, $p \geq 0.077$; $+28.4$/$+28.3$ and $+27.8$/$+36.9$ above the same-seed twins, $p \leq 5.6\times10^{-14}$; accuracy 85.9–88.1%). Pooled over the three seeds with the question-clustered sign-flip test, the arms sit -0.3 and $+3.5$ points from Base (p=0.96, 0.21) and $+24.5$ and $+32.9$ above their twins ($p=1.0\times10^{-5}$).

Reading. The bf16 cap-8192 arms were meant to test whether training-time length pressure drives the erosion, and they do not test it: the configuration the long budget forced also prevented the policy from moving, so their Base-level R and Base-level answer position are what an untrained checkpoint gives. The fp32 arms do test it, in three seeds, and the answer is that removing the cap removes the erosion: a policy that demonstrably trained for 200 steps at the registered precision, data order and step budget, differing from its registered twin in the cap and in the memory engineering the cap forces, ends 28–37 points above that twin and at Base in every seed. The remaining micro-batching and memory-scheduling changes preserve the nominal objective and optimizer-step configuration, but their effects under effective fp32 training are not isolated by the near-frozen bf16 control. Training-time length pressure is therefore supported as the mediator by a manipulation, not only by the observational record (every registered arm that trained did so under a cap that zero-rewarded 17–73% of each batch and eroded; the public length-penalty checkpoints erode; the public RL-only checkpoints trained at long budgets move the other way, Appendix G). Two qualifications bound the claim. The support is three training seeds of one base, where the observational side rests on five 1.5B arms and two families. And the manipulation does not isolate length pressure from learning as such: with the cap gone, three quarters of the groups carry no advantage, so the arm shows that the recipe without length pressure neither learns much nor erodes, not that a recipe that learned as much by other means would keep its recoverability; the public RL-only pairs are the only evidence on that point. Two lessons are methodological. A control that only matches the manipulation's engineering can expose an invalid manipulation that its own three-seed replication could not, because replication under a frozen optimizer replicates the freeze; and a GRPO arm's flat reward and length are not by themselves evidence that it did not train, the weight diff and the zero-advantage share have to be read together, since a policy with nothing to learn looks like a policy that cannot. The representational mechanism, why length pressure, where it acts, relocates the moment of computation, remains open.

The cap at 8B. The same manipulation from the Llama-8B base, in the registered bf16 Llama recipe of Appendix A (`--bf16-weights`, cap 8,192 instead of 1,280, per-device batch 1, vLLM utilization 0.35 after a probe ladder, 200 steps in 965 minutes on the A100, one checkpoint at step 200, hashed, evaluated through the Llama base-grid protocol and deleted for disk; runner `run_cap8192_llama8b_arm.sh`, report `paper/cap_llama8b_results.json`), reproduces the bf16 pattern rather than the fp32 one. R is 74.2% (196/264) against the base's 74.3% (-1.2, p=0.80) and $+15.7$/$+16.9$ points above the registered 1,280-cap α=0 arms of seeds 7 and 17 at step 200 (57:17, $p=3.4\times10^{-6}$; 62:19, $p=1.8\times10^{-6}$), accuracy 88.0%; but the step-200 weights differ from the base in 4.6% of parameters (relative change 2.6×10^{-5}), with 80% of prompt groups at zero advantage against 31% in the 1,280-cap arm, 2.5% of logged steps touching the cap and a median batch length of 2,298 tokens. Under bf16 the long cap again left the optimizer little to register, so the 8B arm is consistent with the cap account and does not test it; an fp32 8B run does not fit the rented card, and the mediator claim stays scoped to the three 1.5B seeds.

E MATH-500 1.5B CONDITIONAL SURVIVAL

Table 13 complements the GSM8K survival table (Table 5) with the three 1.5B grids on MATH-500. The compressor ordering observed on GSM8K largely carries over for Base and L1-Max, while ThinkPrune's cells are uniformly the weakest; StreamingLLM remains comparatively strong for L1-Max (79.9% at 75%).

Table 13: MATH-500 conditional answer survival (%) on each 1.5B model's eligible set.

Model	Compressor	.25	.50	.75	.90
Base	SnapKV	95.6	89.4	76.9	54.4
Base	TOVA	97.5	86.2	70.0	45.6
Base	StreamingLLM	90.0	86.9	75.0	53.8
L1-Max	SnapKV	92.8	83.5	71.9	56.1
L1-Max	TOVA	95.7	89.9	66.9	43.9
L1-Max	StreamingLLM	84.9	84.2	79.9	57.6
ThinkPrune	SnapKV	90.4	80.9	62.8	40.4
ThinkPrune	TOVA	87.2	80.9	61.7	37.2
ThinkPrune	StreamingLLM	79.8	70.2	60.6	35.1

F THE CROSS-MODEL CONTRAST ON A NON-MATHEMATICAL TASK: MMLU-PRO

The registered cross-model contrast is measured on GSM8K and MATH-500, and both length-penalty checkpoints were trained with *mathematical* outcome rewards, so the deficit reported in Section 4.1 could be an artifact of evaluating them on their own reward distribution. This appendix asks whether it persists off that distribution by running the identical `midpoint-v1` protocol on MMLU-Pro for the same 1.5B trio (`run_mmlu_trio.sh`; three 300-row grids, registered as `valid` in the manifest; report `paper/mmlu_trio_results.json`, generated by `src/analyze_mmlu_trio.py` through the registered estimands). The analysis is *post-hoc and exploratory*: it was specified and run after the confirmatory grids, and three protocol choices deliberately differ from them, a 400-row source pool instead of 320 (the 1.5B models are much weaker on MMLU-Pro; the grid still takes the first 300 preflight-valid rows), bfloat16 generation for all three pools (the registered GSM8K L1-Max and ThinkPrune pools were fp16, and the registered BF16 regeneration of that pool shifted R nonsignificantly, Section 3.4, so a uniform dtype removes that confound on a new task), and L1-Max keeping its published length instruction, as in the MATH-500 extension. Everything else is unchanged: temperature 0.6 / top-p 0.95 / seed 7, a 4096-token budget, the answer-marker preflight at f=0.5, and SnapKV/TOVA/StreamingLLM at $r \in \{0.25.5.75.9\}$ with the same 64-token forced elicitation. Because MMLU-Pro gold answers are option letters A–J, the gold-string presence audit of Appendix K does not apply here.

The deficit persists, and does not shrink. $R(0.5)$ is 67.7% for Base (63/93 correct sources; Wilson 95% CI 57.7–76.4), against 36.4% for L1-Max (39/107; 27.9–45.9) and 24.2% for ThinkPrune (31/128; 17.6–32.3); the already registered DeepSeek-7B MMLU-Pro grid sits at 64.1% (100/156) through the same code path. Pairing by question text on jointly source-correct questions, Base leads L1-Max by $+42.1$ points (26 versus 2 discordants, n=57, two-sided exact McNemar $p = 3.0\times10^{-6}$) and ThinkPrune by $+45.7$ points (35 versus 3, n=70, $p = 6.7\times10^{-8}$). Recomputing the same contrasts from the registered GSM8K trio grids gives $+40.0$ points (n=250, $p = 1.9\times10^{-22}$) and $+37.1$ points (n=245, $p = 1.4\times10^{-20}$): the gap on the non-mathematical task is if anything *wider* (by 2.1 and 8.6 points) than on GSM8K, on far fewer paired questions. The erosion is therefore not domain-specific to mathematics, and we read it as a property of the checkpoints' mid-trace state rather than of the reward's task distribution, a single non-mathematical task, so this is one replication and not a demonstration of generality.

The gap is not a guessing artifact. On the ten-way multiple choice of MMLU-Pro, blind chance is 10%, so the wrong-source floor of Appendix I is the reference against which every rate must be read. Among *incorrect*-source rows the uncompressed midpoint still elicits the gold option in 24.6% (51/207) for Base, 16.9% (29/172) for ThinkPrune and 5.2% (10/193) for L1-Max, against 28.5% for DeepSeek-7B; under compression these fall to 6.3–9.2%, 2.3–10.5% and 1.6–2.1% at r=0.9, i.e. to the ten-way chance level in eight of those nine cells and just above it in the ninth. The floor thus *orders the models the same way R does*, the RL checkpoints extract less from their own midpoints, correct or not, so the deficit cannot be explained by Base guessing more successfully, and the eligible-set survival rates of Table 14 sit above the floor everywhere except L1-Max's SnapKV and StreamingLLM 90% cells.

Table 14 reports conditional survival on each model's own eligible set. Base is above both RL checkpoints in ten of the twelve cells, the two exceptions being the r=0.9 TOVA and StreamingLLM cells, where ThinkPrune's small eligible set sits nominally higher (31.7% and 25.0% against 22.8% and

21.9%); the GSM8K compressor ordering does not carry over, and as on the 7B MMLU-Pro grid StreamingLLM is the weakest compressor at low ratios (34.7% for L1-Max already at 25% compression). Mean critical ratios are 0.53–0.60 for Base, 0.40–0.50 for ThinkPrune and 0.18–0.39 for L1-Max, with 0–4 per-record non-monotonicities per cell. The matched analogue of Appendix AE, questions jointly source-correct and jointly eligible for all three models, leaves only n=7 questions here, and every matched McNemar is null ($p \geq 0.25$); that subset is too small to test anything and is reported as underpowered, not as an absence of difference.

Table 14: MMLU-Pro conditional answer survival (%) on each 1.5B model's own eligible set (Base n=114, L1-Max n=49, ThinkPrune n=60 of 300 rows).

Model	Compressor	.25	.50	.75	.90
Base	SnapKV	86.0	71.9	48.2	23.7
Base	TOVA	93.9	86.0	47.4	22.8
Base	StreamingLLM	78.1	71.9	51.8	21.9
L1-Max	SnapKV	69.4	44.9	30.6	6.1
L1-Max	TOVA	71.4	53.1	24.5	14.3
L1-Max	StreamingLLM	34.7	26.5	18.4	2.0
ThinkPrune	SnapKV	75.0	53.3	43.3	16.7
ThinkPrune	TOVA	86.7	61.7	40.0	31.7
ThinkPrune	StreamingLLM	55.0	51.7	45.0	25.0

G PUBLIC RL-ONLY PAIRS AT 7B AND 14B: DEEPSEEK-R1-DISTILL-QWEN VERSUS ACEREASON-NEMOTRON

Every public RL checkpoint in the registered cross-model contrast of Section 4.1 is either a length-penalty recipe (L1-Max, ThinkPrune, L1-Qwen3-8B) or otherwise length-related (DeepScaleR trains by iterative context *lengthening*), so the registered cross-model contrast cannot separate "RL post-training" from "RL post-training that manipulates chain length." This appendix adds the missing cell: a public checkpoint trained by *pure outcome reward* from a base that already has a registered grid. `nvidia/AceReason-Nemotron-7B` (revision `7d385d30`, NVIDIA Open Model License) is, per its model card, "trained entirely through reinforcement learning starting from DeepSeek-R1-Distill-Qwen-7B", the exact base revision (`916b56a4`) of our registered 7B GSM8K grid, with a mathematical outcome reward and no SFT stage, no length reward and no format reward. We ran the identical `midpoint-v1` protocol used for that base grid (`run_remote_queue.sh`: openai/gsm8k test, 320-row pool, 8192-token budget, bfloat16, temperature 0.6 / top-p 0.95 / seed 7, the default instruction, answer-marker preflight at f=0.5 excluding 1 row, SnapKV/TOVA/StreamingLLM at $r \in \{0.25.5.75.9\}$ with 64-token forced elicitation), registered the grid as `valid`, and read the pair through the registered estimands (`src/analyze_public_pair.py`; report `paper/public_pair_acereason7b_results.json`). **The deficit is absent, and runs the other way.** $R(0.5)$ is 96.1% for AceReason-7B (272/283 correct sources; Wilson 95% CI 93.2–97.8) against 85.3% for its base (238/279; 80.7–89.0); on questions paired by text and jointly source-correct, AceReason-7B leads by +10.7 points (39 versus 10 discordants, n=271, two-sided exact McNemar $p = 3.8 \times 10^{-5}$). Matched survival on the jointly source-correct, jointly eligible subset (n=221) points the same way, significantly so for StreamingLLM (+8.1 points at r=0.75, 29/11, $p = 6.4 \times 10^{-3}$; +12.7 at r=0.9, 64/36, $p = 6.6 \times 10^{-3}$) and null for SnapKV ($p = 0.91$ and 0.067) and TOVA ($p = 0.59$ and 0.76). Gold-string presence does not explain it: gold appears in the retained prefix for 72.8% of AceReason-7B's correct sources against 50.2% for the base, yet on the jointly gold-absent, jointly source-correct subset (n=55, frozen rule of Appendix K) AceReason-7B still recovers 52 against 40 (15/3 discordants, $p = 7.5 \times 10^{-3}$). Full-chain accuracy is statistically indistinguishable: 93.7% (299/319) against 92.5% (295/319) on the 319 shared pool questions, 12 versus 8 discordants, $p = 0.50$, the "same score" half of this paper's framing holds for this pair while the mid-trace state differs in the *opposite* direction from every length-penalty arm.

The same comparison at 14B. We repeated the pair one scale up under the identical protocol on a second L40 server: `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B` (revision `1df85071`) against `nvidia/AceReason-Nemotron-14B` (revision `b244ed55`, NVIDIA Open Model License), which its model card likewise describes as trained entirely through reinforcement learning from that base, with math and code outcome rewards, no SFT

stage and no length penalty (320-row pool, 8192-token budget, bfloat16, seed 7, answer-marker preflight at f=0.5 excluding 1 row for the base and 3 for the RL checkpoint, the same three compressors, the same ratios and the same 64-token forced elicitation; report `paper/public_pair_acereason14b_results.json`). **The direction replicates and the gap stays significant, but it is attenuated.** $R(0.5)$ is 93.0% for AceReason-14B (264/284 correct sources; Wilson 95% CI 89.4–95.4) against 85.6% for its base (249/291; 81.1–89.1); paired, AceReason-14B leads by $+6.8$ points (32 versus 13 discordants, n=280, $p = 6.6 \times 10^{-3}$), against $+10.7$ at 7B. Every secondary block weakens to null: matched survival (n=228) favours AceReason-14B in five of six press–ratio cells but no cell is significant (smallest $p = 0.059$, SnapKV at r=0.9, where 7B had two significant StreamingLLM cells), and the jointly gold-absent contrast is $+7.1$ points and nonsignificant (n=99, 15/8 discordants, $p = 0.21$) against $+21.8$ at 7B. Gold presence itself barely moves at this scale (47.5% against the base's 47.4%, where at 7B it was 72.8% against 50.2%), so whatever produces the 14B advantage is not a presence shift. Full-chain accuracy is nearer to detectable than at 7B and leans the other way: the base scores 96.5% (306/317) against 94.0% (298/317) on the 317 shared pool questions (11 versus 3 discordants, $p = 0.057$), so the "same score" premise holds only marginally here, and the recoverability gain is certainly not bought by a better-scoring checkpoint.

Reading. This is post-hoc and observational, one pair per scale on one dataset at two scales, so it establishes no scale generality on its own; the base itself is a distillation of a much larger teacher, and AceReason's data, reward scale, and optimizer are not matched to anything we control. What it does do is rule out the universal reading, that any outcome-RL post-training erodes pre-answer recoverability, and it is why the paper's central claim is stated directionally (post-training *moves* recoverability at an unchanged score) and its erosion claim scoped to the length-penalty and length-related checkpoints and to the tested GRPO recipes. Here outcome-only RL at 7B *improved* it while leaving accuracy alone, which is consistent with two non-exclusive readings we cannot separate with one pair: that the erosion tracks the length-manipulating component of a recipe rather than RL as such (the registered controlled arms already show the penalty deepening the decline at 1.5B, and our own outcome-only α=0 arms still erode R, so recipe details clearly matter beyond the penalty term), and that the effect attenuates with scale, as in the milder 8B cross-sectional gaps, the absence of a detectable arm contrast at 8B (Appendix A) and now the 14B pair's smaller gap and uniformly null secondary blocks, on one pair per scale. We therefore report this pair as a boundary condition on the claim's scope, not as a refutation of the controlled result, and we do not fold it into the confirmatory family. Appendix M supplies the descriptive evidence behind both readings: AceReason-7B is the one checkpoint that moves *earlier* in the trace where the answer is first written (median f_{first} 0.322 against its base's 0.491) and the only one trained at a long completion budget, whereas every arm we control ran under a 1,536/1,280-token cap that truncated and zero-rewarded between a third and two thirds of each on-policy batch, and the training-time test of that mediator we attempted (correctness-only arms retrained at an 8,192-token cap) is uninformative, because its bf16 configuration left the policies essentially untrained (Appendix D).

Table 15: The public RL-only pairs on GSM8K, at 7B and at 14B. Gaps are AceReason minus DeepSeek, so a positive gap means the RL checkpoint is *better*; p is a two-sided exact McNemar test on the paired subset named in each row. Paired subset sizes (7B/14B): recoverability n=271/280, jointly gold-absent 55/99, matched survival 221/228, full-chain accuracy 319/317. In the first row only, the two rates are each grid's own correct-source rate while the gap and p are computed on the paired subset, so the printed rates do not subtract to the printed gap ($+10.7$ and $+6.8$ paired, against $+10.8$ and $+7.4$ unpaired); every other row is paired throughout.

	7B pair				14B pair			
Quantity	DeepSeek	AceReason	gap	p	DeepSeek	AceReason	gap	p
$R(0.5)$, correct sources	85.3%	96.1%	$+10.7$	3.8×10^{-5}	85.6%	93.0%	$+6.8$	6.6×10^{-3}
$R(0.5)$, jointly gold-absent	72.7%	94.5%	$+21.8$	7.5×10^{-3}	78.8%	85.9%	$+7.1$	0.21
Survival, StreamingLLM r=0.75	85.5%	93.7%	$+8.1$	6.4×10^{-3}	82.0%	81.6%	-0.4	1.00
Survival, StreamingLLM r=0.90	59.3%	71.9%	$+12.7$	6.6×10^{-3}	53.1%	60.5%	$+7.5$	0.10
Full-chain accuracy	92.5%	93.7%	$+1.3$	0.50	96.5%	94.0%	-2.5	0.057

H PUBLIC EFFICIENT-REASONING CHECKPOINTS: THREE DOSE LADDERS FROM ONE BASE

The registered cross-model contrast holds one public checkpoint per recipe, so it can show that shortened checkpoints sit below their base but not whether the deficit grows with how hard a recipe shortens. This appendix adds thirteen public checkpoints post-trained from the same registered 1.5B base revision, chosen so that three recipes appear at several declared doses, plus two transfer cells at 7B and 8B. ThinkPrune (Hou et al., 2025) supplies single-stage completion caps of 4,096, 3,072 and 2,048 tokens (the registered ThinkPrune checkpoint is the iterative 4k→3k→2k run, so the 2k dose appears twice, as two separate training runs). The length-penalty sweep of Arora & Zanette (2025) supplies $\alpha \in \{0, 0.05, 0.1, 0.2, 0.4\}$, where α=0 is that recipe's own correctness-only control. Laser-D (Liu et al., 2025c) supplies target lengths of 4,096, 2,048 and 1,024. Two single points complete the 1.5B set: E1-Math (Xu et al., 2026), trained with budget-constrained rollouts, and DLER (Liu et al., 2025b), a truncation-based length penalty. The transfer cells are VeriThinker-7B (Chen et al., 2025), an SFT-only shortening of the registered DeepSeek-7B base, and InfoDensity-8B (Wei et al., 2026a), an information-density RL from the registered Llama-8B base.

Protocol. Every checkpoint ran the registered `midpoint-v1` pipeline unchanged: one 320-row GSM8K test pool (8,192-token budget, bfloat16, temperature 0.6, top-p 0.95, seed 7), the answer-marker preflight at f=0.5 (0 or 1 row excluded per checkpoint), the first 300 valid rows, SnapKV, TOVA and StreamingLLM at $r \in \{0, .25, .5, .75, .9\}$ with the 64-token forced elicitation, and the validator. Revisions are pinned in the report (`paper/public_ext_results.json`, produced by `src/analyze_public_ext.py`). Each checkpoint is paired by question text with the registered Base grid of its family on jointly source-correct questions (two-sided exact McNemar); correlations are Spearman with 20,000 permutations. The runs are post-hoc and observational and are not in the confirmatory family. One deviation: the kvpress SnapKV implementation requires a prefill longer than its 64-token observation window, and the α=0.1, 0.2 and 0.4 checkpoints produce midpoint contexts of 31–57 tokens on 1, 4 and 120 of 300 rows. Those SnapKV cells at nonzero ratios were recorded as skipped (4, 16 and 480 cells) and leave that cell's denominator; ratio-zero cells are never skipped, so R is unaffected. The ThinkPrune-2k pool was regenerated after an out-of-memory abort caused by two concurrent runs, and the α=0.1 grid resumed twice from its written rows under the validator's per-cell recomputation; the run record lists both (`logs/public_ext/DEVIATIONS_lane3.md`).

Every checkpoint sits below its base, and the deficit grows with the dose. Table 16 and Figure 6 give the numbers. The thirteen 1.5B checkpoints sit 19.2–65.7 points below Base ($p \leq 4.3 \times 10^{-7}$), VeriThinker-7B 7.6 below DeepSeek-7B ($p = 0.011$) and InfoDensity-8B 31.1 below Llama-8B ($p = 1.1 \times 10^{-17}$). Within each ladder R falls with the declared dose: cap 4,096 → 3,072 → 2,048 gives 53.8 → 42.4 → 37.6% (the registered iterative 2k run, 39.8%, lands within 2.2 points of the single-stage 2k run); α 0 → 0.05 → 0.1 → 0.2 → 0.4 gives 41.1 → 33.6 → 31.8 → 18.8 → 14.5% (Spearman $\rho = -1$, permutation $p = 0.016$); target 4,096 → 2,048 → 1,024 gives 42.4 → 27.6 → 21.6%. Across all sixteen 1.5B checkpoints R rises with the median realized trace length ($\rho = 0.96$, n=16, $p < 10^{-4}$; 0.96 with Base included), so the three ladders and the single points lie on one curve: recipes that shorten more erode more, and at a similar rate per token whatever the reward form. Full-chain accuracy stays within five points of Base for fifteen of the sixteen and differs detectably only for α=0.2 (-4.7, $p = 0.03$) and α=0.4 (-12.3, $p < 10^{-3}$); the α=0.4 run has collapsed to a median trace of 16 tokens, so it marks where the penalty stops shortening reasoning and starts removing it, while α=0.2 (211 tokens) is the last point with near-Base accuracy. InfoDensity-8B is 4.0 points below its base in accuracy ($p = 0.02$). Matched conditional survival on the jointly source-correct, jointly eligible subsets favours the base in all 90 press–ratio cells (fifteen checkpoints, six cells each), significantly in 78.

Reading. Three points follow. First, the erosion is graded rather than a threshold: declared pressure orders it within a recipe, and realized length orders it across recipes, which is what a length-pressure mediator predicts and what a recipe-specific artifact would not. Second, the public correctness-only run of the Arora–Zanette recipe erodes as much as the penalized runs at its length (41.1%, -36.4), matching our own α=0 arms trained under a completion cap (Section 4.1), so a recipe need not declare a length term to apply length pressure. Third, the SFT-only shortening erodes least: VeriThinker-7B cuts its base's median trace from 1,000 to 622 tokens for -7.6 points, where the

RL recipe with the same absolute median (E1-Math, 622 tokens) loses 30.5; this agrees in direction with the SFT control of Appendix C, though the scales differ (7B against 1.5B), so it is suggestive only. The evidence is observational: one pool per checkpoint, recipes that differ in data, steps and optimizer, single training runs, and Wilson intervals of about ± 6 points on $\sim$260 correct sources, so adjacent ladder steps (α 0.05 against 0.1; the two 2k runs) are not individually distinguishable. The claim is the monotone ordering, not each step.

Table 16: Public efficient-reasoning checkpoints on GSM8K under the registered midpoint-v1 protocol. Accuracy is the full-chain rate of the 300-row grid; $R(0.5)$ conditions on correct sources (numerator/denominator); the gap, discordants (Base only / checkpoint only) and two-sided exact McNemar p are on jointly source-correct questions paired by text with the registered base of the same family (Base 1.5B: 88.0%, R 76.9%, median trace 1,230 tokens; DeepSeek-7B: 93.0%, 85.3%, 1,000; Llama-8B: 92.0%, 74.3%, 1,014). Median length is the pool's median reasoning length in tokens. †: paired full-chain accuracy differs from the base at $p < 0.05$. Skipped: SnapKV cells whose context did not exceed the 64-token window (see text). Rows marked reg. are the registered grids, repeated for reference.

Checkpoint	Recipe / dose	Acc. %	$R(0.5)$ %	Gap	Disc.	p	Med. len.
ThinkPrune	cap 4,096	86.7	53.8 (140/260)	−22.1	72/18	8.1×10^{-9}	684
ThinkPrune	cap 3,072	85.7	42.4 (109/257)	−32.9	88/8	3.7×10^{-18}	608
ThinkPrune	cap 4k→3k→2k (reg.)	86.3	39.8 (103/259)	−37.1	100/9	1.4×10^{-20}	524
ThinkPrune	cap 2,048	87.7	37.6 (99/263)	−38.2	103/9	2.3×10^{-21}	522
Arora–Zanette	α=0	88.3	41.1 (109/265)	−36.4	103/12	3.4×10^{-19}	675
Arora–Zanette	α=0.05	87.3	33.6 (88/262)	−42.7	112/7	1.8×10^{-25}	587
Arora–Zanette	α=0.1	87.0	31.8 (83/261)	−45.1	118/8	3.2×10^{-26}	493 (4 skipped)
Arora–Zanette	α=0.2	83.3†	18.8 (47/250)	−57.8	140/3	8.7×10^{-38}	211 (16 skipped)
Arora–Zanette	α=0.4	75.7†	14.5 (33/227)	−65.7	141/3	4.5×10^{-38}	16 (480 skipped)
Laser-D	target 4,096	87.3	42.4 (111/262)	−34.5	93/7	2.7×10^{-20}	640
Laser-D	target 2,048	87.0	27.6 (72/261)	−49.0	127/6	1.3×10^{-30}	436
Laser-D	target 1,024	86.3	21.6 (56/259)	−55.7	142/6	7.7×10^{-35}	296
E1-Math	budget-constrained rollout	88.7	45.1 (120/266)	−30.5	88/12	1.9×10^{-15}	622
DLER	truncation-based penalty	87.7	56.3 (148/263)	−19.2	67/20	4.3×10^{-7}	834
L1-Max	length instruction (reg.)	89.3	36.2 (97/268)	−40.0	110/10	1.9×10^{-22}	496
DeepScaleR	context lengthening (reg.)	89.0	59.6 (159/267)	−15.8	60/21	1.7×10^{-5}	1,106
VeriThinker-7B	SFT-only, from DeepSeek-7B	95.7	77.4 (222/287)	−7.6	42/21	0.011	622
InfoDensity-8B	information-density RL, from Llama-8B	88.0†	43.2 (114/264)	−31.1	89/9	1.1×10^{-17}	635

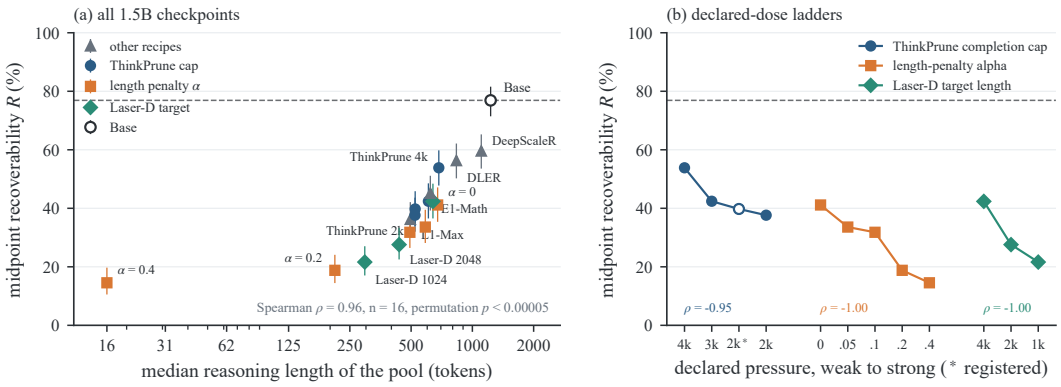


Figure 6: Midpoint recoverability of the sixteen public 1.5B checkpoints post-trained from the registered Base (dashed line). (a) R against the pool's median reasoning length (log scale, Wilson 95% intervals); marker shape gives the recipe family, and ladder interiors are unlabeled. (b) The three declared-dose ladders from weakest to strongest pressure; the hollow marker is the registered iterative ThinkPrune run.

I GOLD RECOVERY FROM INCORRECT SOURCE TRAJECTORIES

The recoverability rates in Section 4.1 condition strictly on correct source trajectories. Table 17 reports the complementary quantity: among trajectories whose final answer is *wrong*, how often the uncompressed midpoint still elicits the gold answer. The rate is highest on the ten-option MMLU-Pro (28.5%), where partial mid-trace knowledge and guessing both contribute, and lowest

for ThinkPrune (4.9–8.2%). For Base on GSM8K, one in four incorrect trajectories already had a midpoint from which the fixed elicitation extracts the gold answer, consistent with late-trace errors, rather than missing extractable intermediate content, causing a visible share of full-chain failures. The same incorrect-source rows give an empirical guessing floor under compression: on MMLU-Pro at r=0.9 they elicit the gold option in only 9.1–13.2% of cases (Wilson 95% CIs within 5–20%), i.e. at the ten-way chance level, whereas eligible-row survival in the same cells of that report is 26–28% (CIs 18–38%): two to three times the floor's point estimate, though the two intervals brush at their nearer ends. (The registered grid puts the same three cells at 25–28%, Table 6; the small difference is the chance-check report's own subsample.) Predicted letters cover all ten options at every ratio (report: `paper/mmlu_pro_chance_results.json`).

Table 17: Incorrect-source trajectories whose midpoint elicits the gold answer.

Grid	incorrect sources	midpoint yields gold	rate (%)
Base GSM8K	36	9	25.0
L1-Max GSM8K	32	5	15.6
ThinkPrune GSM8K	41	2	4.9
7B GSM8K	21	5	23.8
7B MATH-500	87	15	17.2
7B MMLU-Pro	144	41	28.5
Base MATH-500	107	13	12.1
L1-Max MATH-500	103	20	19.4
ThinkPrune MATH-500	97	8	8.2
Llama-8B GSM8K	24	2	8.3

J LENGTH-STRATIFIED RECOVERABILITY

Table 18 gives the length-stratified recoverability behind the confound check in Section 4.1. Strata are quartile edges of the pooled 5–95% trace-length range across the three GSM8K 1.5B grids. RL models have almost no support above 1350 tokens; within the shared support, the deficit persists. Because the within-stratum subsets are not composition-matched, a stricter check restricts to questions where *both* models have in-stratum correct sources and pairs them exactly (`src/analyze_stratum_matched.py`, report `paper/stratum_matched_results.json`): Base versus L1-Max 39/63 versus 28/63 recoverable (discordants 17:6, $p = 0.035$); Base versus ThinkPrune 38/67 versus 25/67 (17:4, $p = 0.0072$). The deficit persists under question-level matching within the shared stratum.

Table 18: Conditional recoverability (%) by trace-length stratum (GSM8K, n in parentheses). Dashes mark strata with fewer than eight correct source trajectories.

Stratum (reasoning tokens)	Base	L1-Max	ThinkPrune
309–830	60.0 (80)	35.5 (228)	38.6 (220)
830–1350	76.3 (76)	71.4 (14)	36.4 (22)
1350–1870	87.0 (46)	—	—
1870–2391	95.7 (23)	—	—

K GOLD-STRING PRESENCE IN THE RETAINED PREFIX

Preflight rejects prefixes containing boxed markers or the phrase "final answer," but a prefix can still contain the computed answer in plain form. For each of the nine grids with string-matchable gold answers (the DeepScaleR and Qwen3-pair grids are covered by an identical-rule extension below) we reconstruct the retained prefix exactly (tokenizer slice of the reasoning body at $\lfloor N/2 \rfloor$ tokens) and search for the gold string with numeric word boundaries (frozen specification and deterministic script: `src/analyze_gold_presence.py`, report `paper/gold_presence_results.json`). Table 19 reports the share of correct source trajectories whose prefix contains the gold string, and conditional recoverability within each subset. MMLU-Pro is excluded because its gold answers are bare option letters, where string matching is meaningless. Two findings matter: RL midpoints contain the completed answer far less often than Base (16–21% versus 52% on GSM8K), and the cross-model recoverability gaps persist within the gold-absent subset with per-question paired significance on GSM8K, and nominally for ThinkPrune

on MATH-500 ($p = 0.031$, short of the corrected threshold; Table 20; all pairing is by question text on jointly source-correct questions).

An extension audit applies the identical frozen rule to the three GSM8K grids above (script `src/analyze_gold_presence_ext.py`; report `paper/gold_presence_ext_results.json`). Gold-present rates: DeepScaleR 31.8%, Qwen3-8B 69.1%, L1-Qwen3-8B 50.2%. On the jointly gold-absent, jointly source-correct subset the Base-versus-DeepScaleR gap survives (n=102, 55 versus 40 recoverable, discordants 27:12, exact $p = 0.024$), so the 17-point DeepScaleR gap is not carried by literal answer presence; the small Qwen3-8B pair contrast is unresolved on this subset (n=67, 58 versus 57, discordants 7:6, $p = 1.0$), an underpowered comparison consistent with the milder 8B gaps.

Separator sensitivity of the frozen rule. The frozen rule matches the bare digit string, so a gold answer of 1,000 or more that the prefix writes with thousands separators ("$200,000") does not count as present. Re-running all twenty-seven grids of this paper under a separator-tolerant variant that also accepts the comma-grouped form moves 58 of 7,024 rows into presence (per-grid rates rise by 0.0–2.9 points, e.g. Base GSM8K 51.5%→53.4%, Llama-8B 51.4%→54.3%, AceReason-7B 72.8%→75.6%) and leaves every contrast of Table 20 at the same verdict ($p \leq 4.3 \times 10^{-6}$ on GSM8K, $p = 0.031$ for ThinkPrune and $p = 1.0$ for L1-Max on MATH-500); the one conclusion that moves is the extension's Base-versus-DeepScaleR contrast, which goes from $p = 0.024$ to $p = 0.058$ and no longer separates at 0.05. The registered audit and its frozen specification are unchanged; the sensitivity is reported, not absorbed (deterministic script `src/analyze_separator_sensitivity.py`, report `paper/separator_sensitivity_results.json`, registered as `separator_sensitivity_v1`). The same audit is what promoted the wrong-answer taxonomy of Appendix L and the answer-position analysis of Appendix M to the separator-tolerant rule v2.

Table 19: Gold-string presence in retained prefixes and conditional recoverability (%) by subset (correct sources only). The parentheses give present/correct-source counts in the second column and *recovered* counts in the last two; each subset's denominator is the present count and its complement.

Grid	gold present	recov. | present	recov. | absent
Base GSM8K	51.5 (136/264)	94.9 (129)	57.8 (74)
L1-Max GSM8K	21.3 (57/268)	64.9 (37)	28.4 (60)
ThinkPrune GSM8K	15.8 (41/259)	65.9 (27)	34.9 (76)
7B GSM8K	50.2 (140/279)	98.6 (138)	71.9 (100)
7B MATH-500	55.4 (118/213)	95.8 (113)	53.7 (51)
Base MATH-500	61.1 (118/193)	90.7 (107)	53.3 (40)
L1-Max MATH-500	46.7 (92/197)	82.6 (76)	41.0 (43)
ThinkPrune MATH-500	32.0 (65/203)	63.1 (41)	32.6 (45)
Llama-8B GSM8K	51.4 (142/276)	93.7 (133)	53.7 (72)

Table 20: Paired recoverability contrast on jointly gold-absent, jointly source-correct questions (two-sided McNemar exact).

Contrast	questions n	Base recov.	other recov.	p
GSM8K: Base vs L1-Max	116	63	26	5.7×10^{-8}
GSM8K: Base vs ThinkPrune	109	62	32	1.4×10^{-6}
MATH-500: Base vs L1-Max	53	27	26	1.0
MATH-500: Base vs ThinkPrune	62	33	23	0.031

L WHERE WRONG MIDPOINT ANSWERS COME FROM

Recoverability counts a midpoint as recovered only when the forced elicitation commits to the gold answer; the complement is a *wrong* committed answer, and this appendix asks, descriptively, where that wrong number came from. For every source-correct row whose uncompressed (r=0) elicitation commits incorrectly, we reconstruct the retained prefix exactly as in Appendix K and locate the predicted string with the separator-tolerant standalone-string rule v2, the frozen rule of Appendix K plus the comma-grouped form of a numeric target of 1,000 or more, so that "200000" also matches "$200,000", reducing to the frozen rule exactly below 1,000, assigning it by priority to the retained prefix, the discarded remainder of the model's own trajectory (including its

post-</think> answer statement), the question text, or none of these (*novel*); the four categories therefore partition each grid's wrong answers, and two non-exclusive flags record the overlap the priority hides. Table 21 reports the composition. Base's GSM8K errors split roughly evenly between numbers already written in the retained prefix (39.3%) and numbers absent from the entire trace (49.2%), while both length-penalty checkpoints shift toward the latter: L1-Max 17.6% versus 68.8% (two-sided Fisher exact against Base, $p = 1.3 \times 10^{-3}$ on the in-prefix share and $p = 8.2 \times 10^{-3}$ on the novel share) and ThinkPrune 25.6% versus 63.5% ($p = 0.068$ and $p = 0.065$, no longer separating at 0.05). DeepScaleR moves the same way without separating ($p \geq 0.17$). On MATH-500 only the Base-versus-ThinkPrune in-prefix share reaches nominal significance ($p = 0.0497$); the other three MATH-500 contrasts are null ($p \geq 0.11$), so the pattern is a GSM8K finding. No wrong answer in any of the nine grids was a bare question quantity absent from the reasoning, though 2–10 predictions per grid do occur in the question text as well as in the trace. These are unpaired comparisons of error *compositions* on unequal numbers of errors, computed after the fact and uncorrected for multiplicity: a mechanism-adjacent description of how two failure modes are mixed, not a mechanism claim. Rule v1 of this taxonomy used the frozen rule alone; because gold answers and normalized predictions carry no thousands separators while the traces write them, it misfiled 8 of the 846 wrong answers, all of them large comma-written numbers, among them the worked example of Appendix AG, as *novel*. Rule v2 repairs that. The repair is small but not cosmetic: it raises L1-Max's in-prefix share from 17.1% to 17.6% and its in-remainder share from 11.2% to 13.5%, and it moves the Base-versus-ThinkPrune contrast from $p = 0.046/0.047$ to $p = 0.068/0.065$, so that contrast no longer separates at 0.05 (Appendix K, separator sensitivity). Deterministic script `src/analyze_wrong_answers.py`; report `paper/wrong_answer_taxonomy_results.json`.

Table 21: Source of the committed wrong answer under the uncompressed (r=0) midpoint elicitation, source-correct rows only. The columns partition each grid's wrong answers by priority (retained prefix, then discarded remainder, then question text); the question-only category is empty in all nine grids and is omitted.

Grid	wrong n	in prefix (%)	in remainder (%)	novel (%)
Base GSM8K	61	39.3	11.5	49.2
L1-Max GSM8K	170	17.6	13.5	68.8
ThinkPrune GSM8K	156	25.6	10.9	63.5
DeepScaleR GSM8K	107	29.0	10.3	60.7
DeepSeek-7B GSM8K	41	19.5	4.9	75.6
Llama-8B GSM8K	71	38.0	21.1	40.8
Base MATH-500	45	37.8	20.0	42.2
L1-Max MATH-500	78	34.6	12.8	52.6
ThinkPrune MATH-500	117	22.2	21.4	56.4

M WHERE IN THE TRACE THE ANSWER IS FIRST COMPUTED, AND THE BUDGET THE ARMS TRAINED UNDER

Recoverability at f=0.5 asks whether the *first half* of a trace suffices; this appendix asks the underlying positional question directly, and then measures the training-time budget the controlled arms actually ran under. Both parts are post-hoc, exploratory, CPU-only, and generate nothing: deterministic script `src/analyze_answer_position.py`, report `paper/answer_position_results.json`, registered as `answer_position_v1`.

Answer position. For every source-correct row of all twenty-two GSM8K grids and step-200 arm evaluations analyzed we tokenize the reasoning body before </think> at the pinned revision without special tokens and locate the *first* standalone occurrence of the gold string with the separator-tolerant rule v2 of Appendix L, the frozen rule of Appendix K plus the comma-grouped form of a numeric gold of 1,000 or more, mapping the character-level match back to a token count through the tokenizer's offset mapping. We report $f_{\text{first}} = \text{index}/N$, the *verification tail* $N - \text{index}$ (the answer-stable tail of Chang et al., 2026), and the reasoning-length distribution. As a gate, the same script recomputes gold-presence in the retained prefix through the registered audit's own code path with the *frozen* rule and asserts it reproduces the published per-grid counts exactly (136/264, 57/268, 41/259, 85/267, 140/279, 206/283, 142/276, 201/291, 147/293), so the rule-v2 upgrade leaves the

registered leakage audit untouched. Neither rule is monotone in prefix length, so $f_{\text{first}} \leq 0.5$ and prefix presence disagree on 23 of 5,951 rows (0–5 per grid); the script asserts every disagreement is the same one-directional truncation artifact, the retained prefix ends at the last gold character while the body continues with a word character, a decimal point, or a further comma group, and counts rather than smooths them.

Post-training moves where the answer is written, in the same recipe-dependent direction as recoverability. Table 22 gives the per-grid statistics. Median f_{first} is 0.453 for Base 1.5B and rises to 0.687 (L1-Max), 0.707 (ThinkPrune) and 0.629 (DeepScaleR); paired by question text on jointly source-correct questions with gold present in both bodies, the shift is $+0.213$, $+0.200$ and $+0.118$ in median (two-sided exact sign test $p \leq 3.1 \times 10^{-10}$; paired sign-flip permutation $p = 10^{-5}$, its floor at 10^5 draws). All thirteen step-200 arm evaluations move the same way: the eight 1.5B GRPO arms by $+0.137$ to $+0.221$ ($p \leq 6.4 \times 10^{-15}$), the length-penalty arm deeper than its correctness-only twin in every 1.5B seed, the SFT control least ($+0.050$, $p = 1.8 \times 10^{-3}$), the Llama-8B arms mildly ($+0.027$ to $+0.061$, $p \leq 1.4 \times 10^{-3}$), mirroring the absence of an arm contrast at 8B. AceReason-7B runs the *other* way: median f_{first} 0.322 against its own base's 0.491, -0.129 paired (n=267, $p = 1.3 \times 10^{-15}$). The verification tail states the same fact in tokens: Base, Llama-8B and Qwen3-8B write 480–762 median tokens after first computing the answer, the length-penalty checkpoints 136–148, the 1.5B arms 134–207, and AceReason-7B 1298. In absolute tokens the first occurrence moves the other way, from Base's median 467 to 316–339 in the length-penalty checkpoints and 376–424 in the 1.5B arms (AceReason-7B: 594): the answer is reached slightly earlier while the trace behind it shrinks far more, so fractional and absolute positions disagree in sign, which reconciles the earlier answer fixation reported under RL (Wei et al., 2026b) with the later fractional position here. Post-training shortens the chain mostly behind the answer, leaving the moment of computation near the end of whatever chain remains, except in the one recipe whose recoverability improves.

Truncation exposure. The controlled arms trained with `--max-completion-length` 1536 at 1.5B and 1280 at 8B (`src/train_length_rl.py`; vLLM `max_seq_len` 2048 and 1792 in the logs). A completion that hits the cap is cut off before its boxed answer and scores reward zero, so the cap is an implicit length penalty even in the α=0 control. The TRL logs make the exposure explicit (Table 23): `completions/max_length` equals the cap in 95–100% of logged steps in every 1.5B run and in 100% of steps in all eight Llama-8B runs, and the logged `clipped_ratio` starts at 0.59–0.73 and never falls below 0.10 (1.5B) or 0.29 (8B), between a third and two thirds of every on-policy batch was truncated and zero-rewarded throughout training. A log-free estimate of the exposure at initialization agrees: 45.2% of the registered 1.5B base pool's 500 completions exceed 1536 tokens (39.9% of correct, 81.2% of incorrect) and 46.9% of the Llama-8B pool's 320 exceed 1280 (42.8% and 89.3%). This is the most parsimonious account of why our correctness-only control also erodes R, and it is consistent with AceReason-7B, RL-only at a long budget, being the one checkpoint that improves it. Read on its own this exposure would be only suggestive, the comparison across recipes is unpaired and nothing here is corrected for multiplicity, and the direct test is a single seed: a correctness-only arm retrained at an 8,192-token cap with the registered fp32 weights trains and shows neither erosion nor the answer-position signature tabulated below, where three bf16 replicates that looked the same had not trained (Appendix D). The exposure statistics therefore remain suggestive on their own, and the cap is supported as the mediator by that one manipulation.

N METHOD-GATE IMPLEMENTATION DETAILS

Allocation formulation. The fixed-budget knapsack of Section 4.4 is $\max_{x_{i,r} \in \{0,1\}} \sum_{i,r} x_{i,r}\, \hat{s}_{i,r}$ s.t. $\sum_{i,r} x_{i,r}\, \mathrm{kv}_{i,r} \leq B$ and $\sum_r x_{i,r} = 1$, where $\hat{s}_{i,r}$ is the policy's out-of-fold predicted success for trajectory i at ratio r, $\mathrm{kv}_{i,r}$ the observed retained KV tokens, and the oracle comparator uses true correctness in place of $\hat{s}$.

Exploratory process-signal family. The third, exploratory feature family extracts 26 schedule-time signals from one midpoint prefill with no gold access: layer-wise attention geometry over layer thirds (normalized entropy, top-32 mass, sink mass, local-window mass, question mass), hidden-state segment convergence (cosine trajectory over eight prefix segments: mean, last, min-

Table 22: Answer position in the reasoning body, source-correct rows only. N = reasoning tokens; "gold" = gold string standalone anywhere in the body; f_{first} and the verification tail are conditioned on that; Δ = median paired difference against the model's own base on jointly source-correct questions with gold in both bodies (two-sided exact sign test).

Grid	N med	gold (%)	f_{first} med [IQR]	tail med	Δ (n)	p
Base 1.5B	1128	98.5	0.453 [0.27, 0.67]	536	—	—
L1-Max 1.5B	483	80.6	0.687 [0.49, 0.87]	148	$+0.213$ (198)	2.8×10^{-20}
ThinkPrune 1.5B	496	99.6	0.707 [0.57, 0.90]	136	$+0.200$ (241)	1.2×10^{-23}
DeepScaleR 1.5B	1023	99.3	0.629 [0.42, 0.87]	328	$+0.118$ (244)	3.1×10^{-10}
DeepSeek-7B	955	99.6	0.491 [0.32, 0.72]	442	—	—
AceReason-7B	2189	98.6	0.322 [0.19, 0.49]	1298	$\mathbf{-0.129}$ (267)	1.3×10^{-15}
Llama-8B	972	99.6	0.444 [0.29, 0.69]	480	—	—
Qwen3-8B	1364	99.7	0.373 [0.23, 0.55]	762	—	—
L1-Qwen3-8B	968	95.9	0.481 [0.36, 0.77]	461	$+0.108$ (276)	2.1×10^{-12}
Controlled arms, step 200 (paired against their own base)						
GRPO 1.5B len. pen.	526–609	98.1–99.3	0.679–0.720	134–173	$+0.164$ to $+0.221$	$\leq 1.1 \times 10^{-19}$
GRPO 1.5B α=0	616–668	98.5–98.9	0.663–0.696	179–207	$+0.137$ to $+0.154$	$\leq 6.4 \times 10^{-15}$
SFT 1.5B control	870	98.4	0.548 [0.37, 0.72]	335	$+0.050$ (237)	1.8×10^{-3}
GRPO 8B (4 runs)	694–752	97.8–98.9	0.558–0.607	274–302	$+0.027$ to $+0.061$	$\leq 1.4 \times 10^{-3}$

Table 23: Truncation exposure of the registered GRPO arms, read from their TRL training logs, plus the log-free exposure of each registered base trajectory pool at the same cap. "at cap" is the share of logged steps whose `completions/max_length` equals the cap. The 1.5B rows also cover the freshly retrained seed-7 arms of Appendix R; the original seed-7 length-penalty run retained no training log, which is why its row counts three arms and the α=0 row four.

Arm family	cap	steps	at cap	`clipped_ratio` first → last (mean)
GRPO 1.5B α=0 (4 runs)	1536	40–50	0.98–1.00	0.59–0.67 → 0.21–0.31 (0.36–0.42)
GRPO 1.5B len. pen. (3 runs)	1536	40	0.95–0.98	0.66–0.70 → 0.17–0.24 (0.35–0.37)
GRPO 8B α=0 (4 runs)	1280	20–40	1.00	0.62–0.73 → 0.42–0.64 (0.53–0.62)
GRPO 8B len. pen. (4 runs)	1280	20–40	1.00	0.62–0.72 → 0.46–0.64 (0.51–0.60)
SFT 1.5B control	—	—	—	not logged (not a GRPO run)
Base pool over-cap share (start-of-training exposure): all / correct / incorrect				
R1-Distill-Qwen-1.5B, n=500	1536			45.2% / 39.9% (174/436) / 81.2% (52/64)
R1-Distill-Llama-8B, n=320	1280			46.9% / 42.8% (125/292) / 89.3% (25/28)

imum, slope), prefix surface redundancy (unique-token ratio, 4-gram repetition), and commitment signals (next-token entropy and top-1 probability, teacher-forced log-probability of the fixed elicitation stub). Under the same out-of-fold logistic evaluation against length-only, 0 of 36 GSM8K development cells improve significantly and twelve are nominally worse (per-cell 95% bootstrap intervals, uncorrected). The strongest single signal, late-layer attention entropy, reaches a direction-corrected univariate AUC of 0.63–0.70 per model (maximum over that model's twelve cells) but does not survive the paired comparison; because this exploratory family showed no development-set improvement, no holdout evaluation was performed (machine-readable report: `paper/process_predictor_dev.json`).

Predictors. Every cell (model, compressor, nonzero ratio) fits a scikit-learn pipeline of `StandardScaler` followed by `LogisticRegression` (L2 penalty, C=1, LBFGS, `max_iter`=2000). Out-of-fold probabilities come from a deterministic five-fold `GroupKFold` whose groups are problem indices, so records derived from the same question never cross a train/test boundary. The length-only baseline uses the formatted midpoint context length; the full model adds the thirteen surface features below.

Surface features. Computed from the question and the exact tokenizer-reconstructed midpoint prefix only (no gold answers, no compression outcomes): `num_tokens`, `char_count`, `line_count`, `sentence_count`, `mean_sentence_chars`, `std_sentence_chars`, `digit_ratio`, `equation_marker_ratio` (share of `= + - * /` characters), `repeat_ratio` (repeated word bigrams), `unique_word_ratio`, `correction_count` (sentences containing *wait/correction/actually/reconsider*), `conclusion_count` (occurrences of *therefore/thus/so the answer/final answer/in conclusion*), and `question_overlap`.

Internal features. F1 scores the committed ratio-zero continuation and the gold answer string under the byte-identical uncompressed midpoint elicitation prompt and emits four values per trajectory: sum and mean token log-probability of each. F1 is oracle-style: gold log-probabilities are unavailable at schedule time. F2 concatenates the final-token last-layer hidden state with the mean over the final 64 context tokens. Both were frozen, with lineage validation against the registered grids, before any predictor was fit.

Uncertainty and acceptance. All intervals are 10,000-draw bootstrap 2.5/97.5 percentiles with fixed seeds: per-cell Brier differences resample samples (paired), and allocation accuracy differences resample evaluation batches of 32 queries. Allocation costs are the observed pre-answer KV token counts quantized to 32-token units. The acceptance rule (Section 4.4) requires the full-feature policy's paired-bootstrap lower bound to be positive against both the length-only policy and the matched-memory uniform frontier in the same cell. Table 24 summarizes every gate outcome. The allocation knapsack runs at $B \in \{25, 50, 75\}\%$ of the matched-memory budget and is scored by realized correctness; the length-only policy beats the matched-memory uniform frontier in thirteen of the 36 cells.

Table 24: Method-gate outcomes (negative result). "Better/worse" counts cells whose paired-bootstrap Brier interval excludes zero vs. length-only; the allocation gate also requires beating the matched-memory uniform frontier.

Feature set	Data	Brier vs. length	Alloc. gate
Surface (13)	GSM8K (4 grids)	0/48 better, 20 worse	0/36 pass
F1 (log-probs)	GSM8K dev	mean -0.0037; 4/48 better	—
F2 (hidden)	GSM8K dev	not selected (mean > 0)	—
F1 (selected)	MATH-500 holdout	0/12 better, 4 worse	1/9 pass
Process (26)	GSM8K dev	0/36 better, 12 nom. worse	not run

Predictor-class check (post-hoc, exploratory). The registered gate fits logistic predictors, so its verdict could reflect the predictor class rather than the features. We therefore reran the identical pipeline, keeping the same grouped folds, the same logistic length-only baseline, and the same Brier screen and allocation gate, replacing the full-feature predictor with gradient-boosted trees (HGB), a random forest (RF), or an MLP, over the surface features alone, surface plus the archived F1/F2 internal features, and surface plus F2 only (F1 is oracle-style and undeployable); Table 25. HGB and the MLP fail everywhere. RF over hidden-state features improves paired-bootstrap Brier in 20–22 of 48 cells with at most one cell worse, reproducibly across two forest seeds (mean gain 0.010 Brier), yet passes the allocation gate in only 2–3 of 36 cells, at different cells for each seed, with lower bounds at 0.005–0.031. Twelve configurations were tried post hoc, so these passes are within multiplicity noise. The reading is therefore two-sided: hidden states carry real per-trajectory signal that a nonlinear predictor can calibrate, but in this regime it does not convert into allocation gains over the length-only policy; the gate's rejection of an adaptive method stands, while the blanket "no feature family predicts tolerance" should be read with this qualification (reports: `paper/nonlinear_gate_results*.json`).

Table 25: Post-hoc predictor-class check over the 48 GSM8K method-gate cells (Brier better/worse vs. length-only; allocation passes of 36). Seed 7 unless marked.

Predictor	Features	better	worse	alloc. pass
Logistic (registered)	surface	0	20	0
HGB	surface	0	27	0
RF	surface	1	14	2
MLP	surface	0	45	0
Logistic	surface+F1+F2	0	12	4
HGB	surface+F1+F2	3	7	0
RF	surface+F1+F2	22	1	2
RF (seed 17)	surface+F1+F2	22	0	3
MLP	surface+F1+F2	0	27	0
HGB	surface+F2	0	12	0
RF	surface+F2	20	1	3
RF (seed 17)	surface+F2	20	1	2

O PROMPT ROBUSTNESS OF THE RATIO-ZERO ELICITATION

Table 26 reports the elicitation-wording check behind Section 4.1: 120 source-correct GSM8K trajectories per 1.5B model, re-elicited under three semantically equivalent prompt wordings (greedy, seed 7). Rates are stable within each model, within-item agreement across all three wordings is 92–96%, and the Base-versus-RL ordering is invariant under every wording (machine-readable report: `paper/prompt_robustness_results.json`). All three wordings share the forced answer prefix `The final answer is \boxed{`. Removing the prefix collapses the measurement rather than perturbing it: under free continuation (same contexts and wordings; greedy, 128 new tokens) the models restart chain-of-thought reasoning instead of stating an answer, producing a boxed answer in only 2 of 1,080 rows across the three models at that budget. The forced prefix is therefore a constitutive component of the elicitation, and R is interpreted conditional on it; Appendix P lifts the generation budget and separates immediate extraction from self-completion.

Table 26: Recoverability under three semantically equivalent elicitation prompts (120 source-correct trajectories per model; all-prompt agreement is the share of items with identical outcomes under all three wordings).

Model	original	equivalent a	equivalent b	all-prompt agreement
Base	76.7%	76.7%	76.7%	95.8%
L1-Max	35.8%	38.3%	36.7%	91.7%
ThinkPrune	38.3%	39.2%	39.2%	95.8%

P FREE-ELICITATION CONTROL: IMMEDIATE EXTRACTION VERSUS SELF-COMPLETION

The prefix-free collapse above reflects the 128-token budget, not an inability to answer: with budgets large enough to finish, all three models produce boxed answers from the same re-encoded midpoint contexts without the forced prefill. We rerun 240 deterministic source-correct midpoints per model (seed 7) under two prefix-free wordings, the registered wording ("state ONLY the final answer") and a continuation wording ("Continue the reasoning above to completion…"), greedy at 1,536 new tokens, plus a 3,072-token continuation rerun as a truncation control (script `src/free_continuation.py`; report `paper/free_continuation_results.json`).

Table 27 reports the outcome, with three findings. (i) The forced prefill is a large *positive* intervention for Base (paired 183 vs. 107 recovered at the 3,072 budget, exact $p = 3.8 \times 10^{-14}$), neutral for L1-Max (90 vs. 97, $p = 0.55$), and *negative* for ThinkPrune (94 vs. 148, $p = 7.7 \times 10^{-8}$). (ii) The cross-model deficit is therefore elicitation-conditioned: at 1,536 tokens the Base-versus-L1-Max gap persists under both prefix-free wordings (paired $p \leq 4.1 \times 10^{-7}$), but at 3,072, where L1-Max's boxed-commitment rate rises from 27% to 97%, the contrast is no longer significant ($p = 0.21$), while ThinkPrune *exceeds* Base under every prefix-free condition ($p \leq 2.7 \times 10^{-4}$). (iii) Base commits in only half of its prefix-free continuations even at 3,072 tokens, so its prefix-free rates are budget-limited lower bounds.

Interpretation: R measures *immediate, forced* extraction from the midpoint state, the low-compute extraction regime in which a compressed cache helps only if an answer is elicitable without regenerating thousands of reasoning tokens, though deployed systems need not use this constrained continuation. The RL deficit in R does not imply the pre-answer information is absent: given budget to re-reason, the trio converges or reverses. The registered claims are scoped to R as defined; this control bounds the stronger information-erosion reading, and the dependence on elicitation is itself checkpoint-heterogeneous.

Q COMPRESSED-MIDPOINT FREE CONTINUATION: THE DEPLOYMENT-RELEVANT LOSS

Appendix P lifts the forced prefill but leaves the cache uncompressed; the registered protocol compresses the cache but forces immediate extraction. This control does both at once, which is what a serving stack does: at each registered midpoint we apply the compressor to the retained KV

Table 27: Free-elicitation control (240 source-correct midpoints per model; recovered counts, rate in %). "Forced" is the registered ratio-zero protocol on the same questions; free modes omit the forced prefill. Commit = boxed-commitment rate at the 3,072 budget.

Model	forced	free-imm. 1536	free-cont. 1536	free-cont. 3072	commit
Base	183 (76.3)	103 (42.9)	95 (39.6)	107 (44.6)	50%
L1-Max	90 (37.5)	45 (18.8)	41 (17.1)	97 (40.4)	97%
ThinkPrune	94 (39.2)	152 (63.3)	159 (66.2)	148 (61.7)	89%

state and then let the model *finish its own reasoning* from the compressed cache, scoring only the final boxed answer. Rows are the same 240 deterministic seed-7 source-correct GSM8K midpoints per 1.5B model as Appendix P, with the same re-encoded context construction, the same prefix-free continuation wording, greedy decoding and a 3,072-token budget, no forced prefill and no injected `</think>`. Each row is evaluated in four cells: `none` at ratio zero (the uncompressed comparator, generated through the identical KVPress stack so that compressed and uncompressed continuations are comparable), StreamingLLM at 0.5 and 0.75, and SnapKV at 0.5, for 960 rows per model (script `src/compressed_continuation.py`; report `paper/compressed_continuation_results.json`). The forced comparator is read from the registered midpoint-v1 grids at the identical press–ratio cells. Post-hoc and exploratory: the protocol was written after the registered grids. *Hardware split, disclosed:* Base and ThinkPrune ran on the L40S (torch 2.5.1+cu124) and L1-Max on an RTX 5060 Ti (torch 2.9.1+cu128); everything the protocol fixes is identical across the three run manifests (cells, prompt, decoding, budget, context limit, dtype, sample size and seed), but contrasts involving L1-Max confound checkpoint with environment.

Table 28 reports the outcome, with three findings. (i) *Uncompressed*, free continuation does not separate Base from L1-Max at all (90 vs. 93 recovered on the 205 shared questions, $p = 0.83$) and puts ThinkPrune *above* Base (89 vs. 125 on 206 shared, $p = 2.2 \times 10^{-4}$), reproducing Appendix P's 3,072-token control through the compression stack. (ii) *Once the cache is compressed, re-reasoning does not rescue the RL checkpoints.* L1-Max falls below Base in every compressed cell, StreamingLLM 0.5: 83 vs. 46 (57/20 discordants, $p = 2.9 \times 10^{-5}$); SnapKV 0.5: 64 vs. 44 (43/23, $p = 0.019$); StreamingLLM 0.75: 27 vs. 11 (25/9, $p = 9.0 \times 10^{-3}$), and on the n=50 jointly free-correct subset its continuations survive 13 vs. 2 at StreamingLLM 0.75 (12/1, $p = 3.4 \times 10^{-3}$) and 34 vs. 24 at 0.5 (15/5, $p = 0.041$). ThinkPrune's 36-question uncompressed lead is entirely erased under compression (81 vs. 80, 61 vs. 72, 30 vs. 29; all $p \geq 0.19$, and every jointly-free-correct survival contrast null). (iii) *Free-continuation survival is far below forced-elicitation survival for every model and every compressed cell*: at StreamingLLM 0.5 / SnapKV 0.5 / StreamingLLM 0.75 the free rates are 53.4/48.5/19.4% (Base), 38.9/38.9/6.5% (L1-Max) and 49.3/46.5/15.3% (ThinkPrune) against forced rates of 93.4/82.5/86.9%, 81.1/85.6/70.0% and 76.6/78.7/60.6%. The paired free-versus-forced McNemar over all 240 rows favors forced in every Base cell and every compressed L1-Max cell ($p \leq 4.0 \times 10^{-4}$), and for ThinkPrune only at StreamingLLM 0.75 ($p = 9.8 \times 10^{-5}$), where its large *uncompressed* free advantage (144 vs. 94, $p = 9.7 \times 10^{-7}$) has reversed. The two survival columns condition on different subsets of the same draw (free-correct at ratio zero, n=103/108/144, against forced-correct at ratio zero, n=183/90/94), so they are read as rates on one sample rather than paired item by item.

Interpretation: the immediate-extraction deficit is *not* rescued by letting the model re-reason once the cache is compressed. The elicitation-conditioned convergence of Appendix P is a ratio-zero phenomenon that does not survive eviction: at ratio zero the RL checkpoints match or beat Base given budget to finish, and under any of the three compressed cells they do not. And the probe *under*-states the end-to-end loss, in every compressed cell an answer is far more often recoverable by immediate forced extraction than by letting the request finish from the same compressed cache, so R is an optimistic, not a pessimistic, diagnostic of what a serving stack loses to mid-trace eviction. Scope: one dataset, one model scale, three cells, and the hardware split above.

Table 28: Compressed-midpoint free continuation (compcont-v1; 240 source-correct midpoints per model, greedy, 3,072 new tokens, no forced prefill). "Free correct" is unconditional over that model's own 240-row draw. "Free surv." conditions on the rows whose *uncompressed* free continuation is correct; "forced surv." is the registered forced protocol conditioned on its own ratio-zero-correct rows of the same draw, different denominators, so the two columns are rates on one sample, not a paired test. The last column is the two-sided exact McNemar of unconditional free correctness against Base on questions shared by the two draws (205 for L1-Max, 206 for ThinkPrune). Post-hoc, exploratory; L1-Max ran on different hardware (text). Report: `paper/compressed_continuation_results.json`.

Model	Cell	free correct	free surv.	forced surv.	p vs. Base
Base	uncompressed	103 (42.9)	—	—	—
	StreamingLLM 0.5	93 (38.8)	55/103 (53.4)	171/183 (93.4)	—
	SnapKV 0.5	71 (29.6)	50/103 (48.5)	151/183 (82.5)	—
	StreamingLLM 0.75	32 (13.3)	20/103 (19.4)	159/183 (86.9)	—
L1-Max	uncompressed	108 (45.0)	—	—	0.83
	StreamingLLM 0.5	50 (20.8)	42/108 (38.9)	73/90 (81.1)	2.9×10^{-5}
	SnapKV 0.5	51 (21.2)	42/108 (38.9)	77/90 (85.6)	0.019
	StreamingLLM 0.75	13 (5.4)	7/108 (6.5)	63/90 (70.0)	9.0×10^{-3}
ThinkPrune	uncompressed	144 (60.0)	—	—	2.2×10^{-4}
	StreamingLLM 0.5	93 (38.8)	71/144 (49.3)	72/94 (76.6)	1.00
	SnapKV 0.5	81 (33.8)	67/144 (46.5)	74/94 (78.7)	0.19
	StreamingLLM 0.75	35 (14.6)	22/144 (15.3)	57/94 (60.6)	1.00

R Deployment-style early exit: the forced prefill is the stronger operator

The free-elicitation control keeps the re-encoded user-context construction; a deployment-style *early exit* instead operates on the generation stream. We rebuild each registered midpoint inside the assistant's think block (user turn identical to trajectory generation), inject `</think>`, and let the model answer greedily within 256 new tokens, both bare and with an appended `Final Answer:` hint (the s1-style budget-forcing operator), on the same seed-7 240-row source-correct samples as Appendix P (script `src/eval_early_exit.py`; report `paper/early_exit_results.json`; specification frozen before any run).

Table 29 reports the outcome, with three findings. (i) For the RL checkpoints, think-close early exit reproduces the registered forced rates: paired forced-versus-early-exit differences are nonsignificant in all four RL cells ($p \geq 0.13$), with hint-mode rates of 38.8%/40.0% against forced 37.5%/39.2% on the same rows, so for these models the registered R is the value a deployed early exit attains, not a probe artifact. (ii) For Base the forced prefill is again a large positive intervention ($p \leq 1.7 \times 10^{-29}$): Base commits in only 35–38% of early-exit continuations even with the hint, yet is correct in 68–72% of those commitments, the highest conditional quality of the trio under every elicitation. (iii) Cross-model orderings therefore reverse under early exit (RL above Base, $p \leq 6.1 \times 10^{-3}$ in hint mode), driven by commitment behavior rather than content, consistent with Appendix P. Practically, the forced prefill is itself a deployable zero-cost operator that dominates think-close forcing for Base (76.3% vs. 25.8%) and matches it for the RL checkpoints; under that stronger operator the cross-model contrast is the one the paper measures. Freshly retrained seed-7 GRPO arms from the same base (weights hashed, then deleted) reproduce both patterns: their forced R lands within the published seed ranges (penalty 41.4% inside 35.3–42.3; control 49.6% at the top of 45.0–49.6; penalty below control), and their early-exit rates (30.5–40.0% at 57–77% commitment) sit near their forced rates rather than showing Base's fifty-point split, so the causal arms behave like the public RL checkpoints under the deployed operator.

S Fixed-token-budget early exit: fraction, not token count, is the comparable unit

The midpoint protocol and Appendix R cut each trace at a per-trace *fraction*, which an operator cannot know online; a serving stack enforces a fixed *token* budget. This variant cuts every source-

Table 29: Deployment-style early exit (240 source-correct midpoints per model; correct in %). "Forced" is the registered ratio-zero protocol on the same rows; bare/hint inject </think> without/with a Final Answer: hint, greedy, 256 new tokens. Commit = boxed-commitment rate.

Model	forced	bare	hint	commit (bare)	commit (hint)
Base	76.3	25.4	25.8	35.4%	37.9%
L1-Max	37.5	34.6	38.8	56.7%	98.8%
ThinkPrune	39.2	45.0	40.0	65.4%	99.2%

correct trace at $B \in \{256, 512, 564, 1024\}$ reasoning tokens (564 is half of Base's median reasoning length, the matched-fraction-of-base cut; 564 and 1024 were added post hoc on 2026-09-08 under the identical protocol, report paper/fixed_budget_matched_results.json), closes the think block in place, and compares three exit operators on identical rows (greedy, 256 new tokens; rows whose trace does not exceed B or whose prefix fails the answer-marker audit are excluded and counted; single RTX 5060 Ti, BF16). Table 30 reports the outcome.

Two findings. (i) **The Base-versus-RL sign flips at matched token budgets, exactly as the fraction curve predicts.** A fixed budget lands at very different points of each model's trace: B=256 is median fraction 0.23 of a Base trace but 0.51–0.52, the midpoint, of the shortened RL traces, and the measured rates line up with the registered fraction curve (L1-Max forced 38.9% at its median midpoint vs. 37.5% registered; Base 27.7% at fraction $\sim$0.23, bracketing its registered $R(0.25) = 37.9\%$ from below on the harder >256-token subset). Consequently every paired Base-versus-RL contrast reverses sign (RL higher, $p \leq 2.3 \times 10^{-4}$ at B=256 forced): at equal token budgets the shortened models are simply further through their reasoning. Fixed-budget comparisons therefore conflate trace length with per-fraction recoverability. The deficit of Section 4.1 is a fraction-matched quantity, and the two units must not be interchanged when comparing checkpoints. The matched-fraction-of-base cut makes the same point at the token count where Base is at its own midpoint: at B=564 Base recovers 78.5% while L1-Max and ThinkPrune, already at 0.84 and 0.79 of their traces, recover 89.3% and 78.4%, and on shared questions the shortened models are ahead (forced 88.9% vs. 73.6%, n=72, p=0.0074; 82.1% vs. 60.3%, n=78, p=0.0033); at B=1024 only 4 and 11 RL traces exceed the budget, so that row carries no contrast. (ii) **The prefill operator extends to fixed-budget triggers.** The forced prefill is never significantly worse than either think-close operator (worst -4.7 points, $p \geq 0.075$ across the six contrasts in which it is behind) and is decisively stronger wherever commitment fails: at B=512 it beats bare/hint by $+51.8$/$+46.2$ points on Base ($p \leq 1.2 \times 10^{-31}$) and beats bare by $+18.9$/$+13.5$ on L1-Max/ThinkPrune ($p \leq 4.1 \times 10^{-3}$). Machine-readable report: paper/fixed_budget_results.json.

Table 30: Fixed-token-budget early exit on GSM8K source-correct rows (correct in %). "n.b." counts source-correct rows whose trace does not exceed B (excluded; 0–10 further rows per cell fail the answer-marker audit, counted in the run manifests); frac. is the median cut fraction B/trace length.

Model	B	n	n.b.	frac.	forced	bare	hint
Base	256	264	0	0.23	27.7	28.0	20.8
Base	512	251	11	0.43	72.5	20.7	26.3
Base	564	246	17	0.46	78.5	23.2	25.6
Base	1024	145	118	0.60	94.5	24.1	32.4
L1-Max	256	257	11	0.52	38.9	34.2	40.5
L1-Max	512	111	156	0.81	86.5	67.6	88.3
L1-Max	564	84	183	0.84	89.3	69.0	92.9
L1-Max	1024	4	264	0.96	100.0	75.0	100.0
ThinkPrune	256	255	4	0.51	40.4	45.1	44.3
ThinkPrune	512	111	138	0.76	76.6	63.1	78.4
ThinkPrune	564	88	165	0.79	78.4	65.9	78.4
ThinkPrune	1024	11	248	0.80	81.8	54.5	81.8

T ORACLE-BOUNDARY MEASUREMENTS VERSUS DEPLOYABLE PROCEDURES

Two of our quantities need information an online scheduler does not have, and one does not. We separate them here because the serving reading depends on the distinction.

The boundary fraction is a share of the *finished* trace's length. A scheduler evicting at token 600 does not know whether that is 25% or 75% of what the model will eventually write, so $R(f)$ and the KV-matched frontier of Section 4.3 and Appendix Y are *oracle-boundary* measurements. They compare resource axes and characterize checkpoints; they are not policies, and Section 4.3 labels the frontier as such. The point is not hypothetical: at matched fixed token budgets, which a scheduler can apply, every cross-model ordering reverses (Appendix S), so no serving rule should be read off R directly.

What is deployable is the grouped calibration of Section 4.5. It selects one ratio per (model, task, compressor) cell from held-out conditional survival, needs no per-request prediction and no knowledge of any trace's final length, and is the only procedure we propose. Its guarantee is also the narrower one: coverage of conditional survival on the eligible set, not unconditional accuracy (Appendix AB reports the unconditional numbers separately and claims no comparison).

R enters the serving story only as a diagnostic. It does not set a budget; it explains why the 1.5B RL cells calibrate to no compression at all, which is a fact about those checkpoints that a benchmark score does not expose.

U Prefix negative controls: recoverability comes from prefix content

Recoverability could in principle mix three sources: information in the retained prefix, the model re-solving the question from the prompt alone, and the elicitation template. On every correct-source GSM8K question of the 1.5B trio (n=264/268/259) we evaluate the true retained prefix (*own*) and four control constructions through one shared prefill-and-elicitation stack: the bare question; a length-matched prefix of uniformly sampled non-special vocabulary tokens; retained prefixes of K=3 *other* questions, each truncated to the own-prefix token length (the primary estimate uses only strictly length-matched partner slots: six partner slots per model were shorter than the target, 6/792, 6/804, and 6/777, causing one question per model to be excluded); and the own prefix with its lines shuffled (every row was segmentable; none excluded). The paired baseline is *own* within this stack; the registered grid's ratio-zero outcome serves as a lineage check only, since the control stack is a separate implementation (per-item agreement 90.9%/89.2%/85.3%; rates 76.1%/33.6%/39.0% against registered 76.9%/36.2%/39.8%). The two stacks are consistent at the rate level, not claimed equivalent per item.

Table 31 reports the outcome. Question-only recoverability is 4–6% for all three models: the fixed elicitation alone recovers almost nothing, and the true prefix adds $+72.0$/$+28.0$/$+34.4$ points (paired McNemar $p \leq 4 \times 10^{-19}$). Random and other-question prefixes sit at or below the question-only floor (cluster-bootstrap own-minus-other gaps 0.753/0.323/0.371, all 95% CIs excluding zero), so the gain is specific to the model's own reasoning content, not to context length. Question-only rates are similarly low across models, providing no evidence that differential question-only re-solving accounts for the Base-versus-RL gap; the much larger own-prefix gains instead show that the retained prefixes carry model-dependent extractable information under this elicitation. Finally, shuffling the prefix lines costs Base 5.3 points ($p = 0.013$) while producing no detectable change for L1-Max or ThinkPrune ($p \geq 0.25$): much of the extractable signal survives coarse segment reordering. The same controls on DeepSeek-7B/MMLU-Pro (n=156) give question-only recovery of 28.2%, the multiple-choice knowledge-plus-guessing baseline that matches the 28.5% incorrect-source rate of Appendix I, against 63.5% with the model's own prefix ($+35.3$ points), with other-question prefixes at 9.9%. This is consistent with concurrent evidence that reasoning models can be insensitive to substantial step removal or reordering (Chen et al., 2026), though our test operates on an audited pre-answer prefix with fresh elicitation. These null results do not establish order invariance, and we draw no mechanistic conclusion (machine-readable report: `paper/prefix_controls_results.json`).

V Role-preserving replay validation of the re-encoded context

The registered protocol re-encodes the question and retained prefix as fresh text (Section 3.1); a model generating online instead conditions on the templated user turn followed by its own reasoning tokens, and teacher-forcing that exact token sequence reproduces the rolling KV state at the

Table 31: Ratio-zero recoverability of the true prefix versus negative controls, all through one shared stack on correct-source questions (GSM8K 1.5B trio; DeepSeek-7B/MMLU-Pro added 2026-08). "Gain" is own minus question-only (paired); other-question rates average up to $K=3$ strictly length-matched partners, and questions with no matched partner are excluded.

Model	own	question-only	random	other-question	shuffled	gain (pts)
Base	76.1%	4.2%	1.5%	0.8%	70.8%	+72.0
L1-Max	33.6%	5.6%	4.5%	1.4%	33.2%	+28.0
ThinkPrune	39.0%	4.6%	1.9%	1.7%	35.9%	+34.4
7B / MMLU-Pro	63.5%	28.2%	25.0%	9.9%	61.5%	+35.3

midpoint under deterministic decoding. To test whether re-encoding changes the measured quantity, we evaluate both constructions through one shared prefill-and-elicitation implementation on 300 GSM8K trajectories per 1.5B model ($f=0.5$; greedy; SnapKV, TOVA, and StreamingLLM at every nonzero grid ratio, 0.25–0.9). Table 32 reports per-item agreement of the correctness indicator: 87–93% at ratio zero and 91.5–92.2% over the 3,600 compressed cells per model (89–96% within every individual ratio). Base aggregate counts are nearly identical under either construction (211 versus 211 at ratio zero); L1-Max recovers *more* under replay (108 versus 93 at ratio zero; 912 versus 824 compressed) while ThinkPrune is essentially unchanged (104 versus 106; 814 versus 785), so the registered re-encoded protocol slightly overstates the Base-versus-RL gap magnitude while preserving its presence and the cross-model ordering (211 versus 108 and 104 of 300). These counts are not conditioned on source correctness, so they are not comparable to R in the main text; the next paragraph computes R itself under both constructions (machine-readable report: `paper/replay_consistency_results.json`).

Table 32: Role-preserving replay versus the registered re-encoded context: correct counts under each construction and per-item agreement of the correctness indicator (300 trajectories per model; compressed columns aggregate the twelve press×ratio cells, $n=3600$).

	Ratio zero ($n=300$)			Compressed cells ($n=3600$)		
Model	replay	re-enc.	agreement	replay	re-enc.	agreement
Base	211	211	92.7%	1934	1936	92.2%
L1-Max	108	93	87.0%	912	824	91.7%
ThinkPrune	104	106	87.3%	814	785	91.5%

The headline quantities measured directly on the rolling state. Agreement measures faithfulness but never states R itself in the rolling state, so we recompute Eq. 1 and both paired contrasts from the same records, now conditioned strictly on a correct source trajectory as in the main text (Table 33; report `paper/rolling_cache_results.json`, script `src/analyze_rolling_cache.py`; no new generation). On the rolling KV state Base recovers 77.3% of audited midpoints against 39.6% for L1-Max and 39.4% for ThinkPrune, matching the registered re-encoded grid (76.9/36.2/39.8%). The paired contrasts survive unchanged in magnitude and significance: Base versus L1-Max +37.6 points (99/5 discordants, $n=250$, $p = 9.5 \times 10^{-24}$) and Base versus ThinkPrune +38.8 points (103/8, $n=245$, $p = 3.7 \times 10^{-22}$), against +42.0 and +37.6 points under re-encoding on the identical questions. A within-model paired test localizes the entire construction effect: re-encoding costs L1-Max 6.0 points (26 versus 10 discordants, $p = 0.011$) and is undetectable for Base (11 versus 8, $p = 0.65$) and ThinkPrune (18 versus 17, $p = 1.0$). Conditional survival is preserved as well, including the compressor ordering: on the rolling state StreamingLLM again leads all three models at 75% compression (83.4/64.8/53.8% versus 69.2/58.3/51.9% for SnapKV and 62.6/48.1/46.2% for TOVA). The main-text finding is therefore not an artifact of re-encoding; the overstatement re-encoding introduces is real but confined to L1-Max, smaller than the deficit by a factor of six, and already credited against the reported gaps in Section 4.1.

W BOUNDARY-FRACTION PAIRED CONTRASTS

Table 34 tabulates the statistics behind Figure 3. All contrasts pair by question text and condition on the source trajectory being correct on every side. Between adjacent Base fractions, 106 questions become recoverable and 4 are lost from $f=0.25$ to $f=0.5$ (McNemar exact $p = 9.2 \times 10^{-27}$), and

Table 33: Midpoint recoverability $R(0.5)$ on the rolling KV state, conditioned strictly on a correct source trajectory, beside the re-encoded arm of the same records and the registered grid. The last column is a within-model two-sided exact McNemar test of the two constructions on the same trajectories.

Model	n src. correct	rolling	re-encoded	registered	constr. disc. (p)
Base	264	77.3%	76.1%	76.9%	11/8 (0.65)
L1-Max	268	39.6%	33.6%	36.2%	26/10 (0.011)
ThinkPrune	259	39.4%	39.0%	39.8%	18/17 (1.0)

57 versus 1 from f=0.5 to f=0.75 ($p = 4.1 \times 10^{-16}$). Cross-model paired McNemar tests versus Base are significant at every fraction ($p \leq 2.8 \times 10^{-7}$ for all six comparisons).

Table 34: Recoverability $R(f)$ by boundary fraction and marginal $\Delta R = R(0.75) - R(0.25)$, both on each model's own correct-source set ($n = 264/268/259$); the slope p comes from the per-question paired exact test on the jointly source-correct subset ($n = 249/245$; $\alpha = 0.025$ after the two-comparison correction). On those paired subsets ΔR is 61.0 (Base) vs. 54.2 (L1-Max) and 61.6 (Base) vs. 73.9 (ThinkPrune); Section 4.3 quotes the second pair, which is the one the slope test compares.

Model	$R(0.25)$	$R(0.50)$	$R(0.75)$	ΔR (pts)	slope vs. Base p
Base	37.9%	76.9%	98.1%	60.2	—
L1-Max	17.5%	36.2%	71.3%	53.7	0.21
ThinkPrune	15.1%	39.8%	88.0%	73.0	0.0035

X BOUNDARY CURVES FOR THE 7B PUBLIC PAIR

Protocol. The public RL-only pair of Appendix G is measured only at the midpoint, so it cannot say *where* in the trace its advantage lives. This appendix reads it as a curve on the instrument of Appendix W. Each model's own registered 320-row GSM8K trajectory pool, the same pool its f=0.5 grid was cut from, was re-run through the `midpoint-v1` protocol at f=0.25 and f=0.75 with everything else unchanged (`run_boundary_7b_pair.sh`: answer-marker preflight at the new fraction, then SnapKV/TOVA/StreamingLLM at $r \in \{0.25.5.75.9\}$ with the same 64-token forced elicitation, bfloat16, seed 7, 300 rows). No trajectory was regenerated, so questions pair across fractions and across models by text, exactly as in the 1.5B trio. The 75% prefix trips the answer-marker audit more often than the 25% one, as expected, 2 excluded pool rows against 0 for DeepSeek-7B, 4 against 1 for AceReason-7B, but all four grids still reached the full 300 rows. The four grids are registered as `valid`; deterministic script `src/analyze_boundary_7b_pair.py`, report `paper/boundary_7b_pair_results.json`, registered as `boundary_7b_pair_v1`. Like the pair itself, this is *post-hoc and exploratory* and is not folded into the confirmatory family.

Result. AceReason-7B sits above its base at every fraction, and its curve is flatter, not steeper (Table 35, Figure 7). The paired gap is +33.6 points at f=0.25 and +10.7 at f=0.5, and +0.7 at f=0.75 where both models have saturated (97.9% and 98.2%, $p = 0.69$). The direction never reverses, and the two significant gaps are the two boundaries at which either model has room to differ. Within-model transitions locate the difference: the base gains 103 questions and loses 3 from f=0.25 to f=0.5 ($p = 4.9 \times 10^{-27}$) and a further 38 versus 2 to f=0.75 ($p = 1.5 \times 10^{-9}$), whereas AceReason-7B, already at 82.0% by a quarter of its trace, gains only 42 versus 2 ($p = 1.1 \times 10^{-10}$) and then 9 versus 3 ($p = 0.15$, its only null transition). The prespecified slope statistic makes this a contrast rather than a description: $\Delta R = R(0.75) - R(0.25)$ is +15.4 points for AceReason-7B against +48.7 for its base on the n=267 jointly source-correct subset (98 questions where the base gains more against 9 the other way, paired exact $p = 4.9 \times 10^{-20}$). Conditional survival keeps its usual structure at every fraction, and favours the RL checkpoint where the curves separate: at f=0.25, StreamingLLM retains 70.5% of AceReason-7B's eligible answers at r=0.75 against 52.4% for the base, with SnapKV 63.7% against 49.0% and TOVA 49.8% against 42.7%; by f=0.75 the two are close (StreamingLLM 97.8% against 96.1%). On the TRACE-style instrument of Appendix Z the pair therefore *rises*: the normalized AUC over $[0.25, 0.75]$ is 0.934 against 0.794, +14.0 points, 113 questions higher against 17 with 137 tied, at the 10^5-draw permutation floor. We report only that primary statistic here; the full-range sensitivity variant anchors f=0 at each model's *registered* question-only rate, and that control exists only for the 1.5B trio (Appendix U).

Reading. The scope of this pair is unchanged from Appendix G, one observational public pair, one dataset, one scale, chosen after the confirmatory grids, so what the curve adds is a positional statement, not new generality. It is the positional statement Appendix M predicts. There, AceReason-7B is the single checkpoint that writes its answer *earlier* in its own chain than its base (median f_{first} 0.322 against 0.491, -0.129 paired, $p = 1.3 \times 10^{-15}$) while every length-penalty and length-capped arm writes it later; here that front-loading is exactly what a boundary curve that is already high at f=0.25 and flat thereafter looks like, and the length-penalty checkpoints' back-loading is what ThinkPrune's steeper curve looks like. The two measurements are not independent, both read the same trajectories, but they are different operators, one positional and one behavioural, and they agree. The practical consequence is the one the paper states directionally: post-training moves where a trace becomes answerable, and the sign of that move is a property of the recipe. It also bounds the TRACE reading of Appendix Z: a rise on that instrument is obtained here by the one checkpoint whose recoverability *improved* at an unchanged full-chain score, so on our operator the rise and fall mark the direction of an answer-position shift and not, by themselves, reward hacking.

Table 35: Boundary curve for the 7B public pair on GSM8K. $R(f)$ is on each grid's own correct sources with 95% Wilson intervals; the paired gap is AceReason − DeepSeek on questions paired by text and jointly source-correct at that fraction, tested by two-sided exact McNemar. The final block is on the n=267 questions jointly source-correct in both models at both endpoints; its columns are DeepSeek-7B, AceReason-7B and the contrast. ΔR uses the paired per-question exact test of Appendix W; the AUC is the normalized trapezoid over $[0.25, 0.75]$ with the sign-flip permutation of Appendix Z.

	f=0.25	f=0.50	f=0.75
DeepSeek-7B $R(f)$	49.5% [43.6, 55.3]	85.3% [80.7, 89.0]	97.9% [95.4, 99.0]
correct sources n	279	279	280
AceReason-7B $R(f)$	82.0% [77.1, 86.0]	96.1% [93.2, 97.8]	98.2% [95.9, 99.2]
correct sources n	283	283	283
paired gap (pts)	+33.6	+10.7	+0.7
disc. (Ace/base), n	101/10, 271	39/10, 271	4/2, 269
exact McNemar p	4.4×10^{-20}	3.8×10^{-5}	0.69
Endpoint statistics, n=267; columns: DeepSeek-7B, AceReason-7B, contrast			
$\Delta R = R(0.75) - R(0.25)$ (pts)	+48.7	+15.4	98/9, $p = 4.9 \times 10^{-20}$
$\text{AUC}_{[0.25,0.75]}$	0.794	0.934	+14.0 pts, $p \leq 10^{-5}$

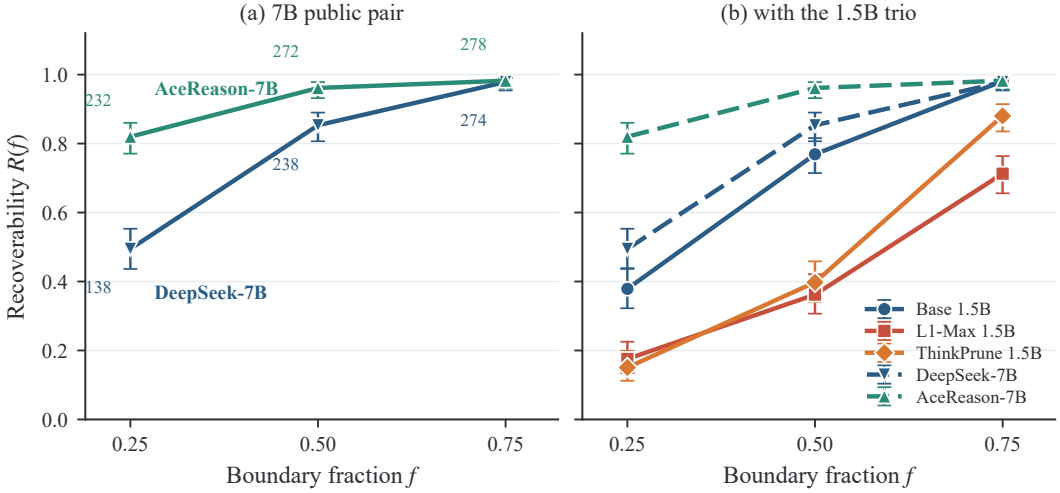


Figure 7: Recoverability $R(f)$ for the 7B public pair, on correct sources with 95% Wilson intervals, in the style of Figure 3. (a) the pair alone, annotated with recoverable counts; the RL-only checkpoint is above its base at every fraction and is already at 82.0% by a quarter of its trace. (b) the same two curves (dashed) against the 1.5B trio (solid) of Figure 3; both base models are navy and take separate markers. The length-penalty checkpoints fall below their base and ThinkPrune steepens, while the RL-only checkpoint rises above its base and flattens.

Y KV-MATCHED TRUNCATION-VERSUS-COMPRESSION FRONTIER

Late truncation plus strong compression wins the KV-matched frontier. Each fraction grid carries the full compressor grid and exact KV accounting, so every (fraction, compressor, ratio) policy is one point in the KV–accuracy plane (Figure 8). This is an *oracle-boundary* measurement, not a deployable policy: fractions truncate at a share of the finished trace's length, which no scheduler knows in advance (Appendix T). The Pareto frontier of all 45 policies per model is dominated by late truncation with strong compression: $f{=}0.75$ with StreamingLLM at 75% compression beats the *uncompressed* midpoint on paired questions with under half the KV: Base 82.6% at 409 vs. 70.6% at 1068 mean KV tokens; L1-Max 57.9% vs. 33.8%; ThinkPrune 73.0% vs. 35.0% (nominal $p \leq 5.6 \times 10^{-5}$; post-hoc cell, surviving Bonferroni over the 45 per-model policies, $p \leq 2.5 \times 10^{-3}$, and all 135 policy–model pairs, $p \leq 7.5 \times 10^{-3}$). The margin is largest for the RL checkpoints, whose weak midpoints are partly rescued by keeping more of the trace (report: `paper/frontier_results.json`).

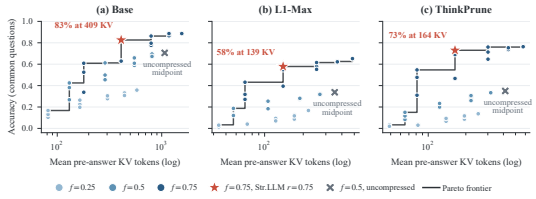


Figure 8: KV-matched truncation-versus-compression frontier on GSM8K (unconditional accuracy over common questions; x = mean pre-answer KV tokens, log). Colors: truncation fraction; black step: Pareto frontier; the starred late-compressed policy dominates the uncompressed midpoint (cross) in every model.

Z TRACE-STYLE TRUNCATION AUC: DIRECTION OF THE EFFECT

TRACE integrates expected reward against truncation fraction into a per-response AUC and flags a *rise* over the pre-RL initial policy as implicit reward hacking, the answer becoming available too early (Wang et al., 2026c). Because our protocol shares the truncate-and-force operator and a pre-RL baseline, we compute the closest quantity our registered grids support rather than assert a direction. This is post-hoc and exploratory.

From the three GSM8K boundary grids per model we form, for each question, the normalized trapezoid AUC over the measured range $[0.25, 0.75]$, which reduces to $0.25\,r(0.25) + 0.50\,r(0.50) + 0.25\,r(0.75)$ with r the per-question recoverable indicator of Appendix W. Questions enter when jointly source-correct in all three models at all three fractions ($n = 239$). A sensitivity AUC over the full $[0, 1]$ range anchors $f{=}0$ at each model's registered question-only rate (Appendix U) and $f{=}1$ at 1.0. Significance uses a two-sided paired sign-flip permutation on per-question AUC differences (10^5 draws, fixed generator seed).

Both RL checkpoints *fall* relative to Base on this instrument (Table 36), by 31.2 and 26.3 points on the primary statistic and 21.3 and 17.2 on the sensitivity variant, all at the permutation floor. The direction is opposite to the rise TRACE reports for hacking policies. Three caveats bound the comparison: three fractions rather than a dense sweep, greedy binary commitment rather than expected reward over five samples at temperature 0.7, and our answer-marker audit and correct-source conditioning, which TRACE does not apply. The two measurements therefore read opposite tails of one operator, and our GSM8K arms carry no reward-model loophole for a hacking signature to exploit. The direction is recipe-dependent, not a property of our protocol: on the same instrument the RL-only public pair *rises*, AceReason-7B scoring 0.934 against its base's 0.794 ($+14.0$ points, 113 questions higher against 17 on $n{=}267$, permutation floor), which is the TRACE-flagged direction obtained here by a checkpoint whose recoverability *improved* (Appendix X).

AA CALIBRATED COMPRESSION BUDGETS: FULL RESULTS

Table 37 lists the calibrated budget and realized held-out KV saving for every grid, compressor, and retention target under the procedure of Section 4.5. All 48 nonzero deployed cells meet their

Table 36: TRACE-style truncation AUC on the 239 jointly source-correct GSM8K questions. Primary is the normalized trapezoid AUC over the measured $[0.25, 0.75]$ range; sensitivity anchors $f{=}0$ at the question-only rate and $f{=}1$ at 1.0. Contrasts are paired per question against Base; p is at the 10^5-draw permutation floor in every cell.

Model	$\text{AUC}_{[0.25,0.75]}$	vs. Base (pts)	disc. (model/Base)	$\text{AUC}_{[0,1]}$	vs. Base (pts)
Base	0.719	—	—	0.659	—
L1-Max	0.407	−31.2	18/153	0.447	−21.3
ThinkPrune	0.456	−26.3	16/133	0.487	−17.2

held-out target (minimum margin +2.1 points; 32/13/3 at targets 75/90/95%); the other 69 fall back to no compression, safe by construction on the eligible set.

Table 37: Calibrated budgets r^* (realized held-out mean KV saving) at retention targets 75/90/95% with $\delta{=}0.05$; "—" prescribes no compression.

Grid	Compressor	target 75%	target 90%	target 95%
Base-1.5B / GSM8K	SnapKV	0.25 (24%)	—	—
	TOVA	0.5 (48%)	0.25 (24%)	—
	Str.LLM	0.75 (72%)	—	—
L1-Max-1.5B / GSM8K	SnapKV	0.25 (23%)	—	—
	TOVA	—	—	—
	Str.LLM	—	—	—
ThinkPrune-1.5B / GSM8K	SnapKV	0.25 (23%)	—	—
	TOVA	0.25 (23%)	—	—
	Str.LLM	0.25 (23%)	—	—
DeepScaleR-1.5B / GSM8K	SnapKV	—	—	—
	TOVA	0.25 (24%)	—	—
	Str.LLM	0.25 (24%)	—	—
DeepSeek-7B / GSM8K	SnapKV	0.5 (48%)	0.25 (24%)	0.25 (24%)
	TOVA	0.5 (48%)	0.25 (24%)	—
	Str.LLM	0.75 (72%)	0.25 (24%)	—
Llama-8B / GSM8K	SnapKV	0.5 (48%)	0.25 (24%)	—
	TOVA	0.5 (48%)	0.25 (24%)	—
	Str.LLM	0.75 (72%)	—	—
Qwen3-8B / GSM8K	SnapKV	0.5 (48%)	0.25 (24%)	0.25 (24%)
	TOVA	0.5 (48%)	0.25 (24%)	—
	Str.LLM	0.75 (72%)	0.5 (48%)	0.5 (48%)
L1-Qwen3-8B / GSM8K	SnapKV	0.5 (47%)	0.25 (24%)	—
	TOVA	0.5 (47%)	—	—
	Str.LLM	0.75 (71%)	0.25 (24%)	—
Base-1.5B / MATH500	SnapKV	0.5 (49%)	—	—
	TOVA	0.5 (49%)	0.25 (24%)	—
	Str.LLM	0.5 (49%)	—	—
L1-Max-1.5B / MATH500	SnapKV	0.25 (24%)	—	—
	TOVA	0.5 (48%)	—	—
	Str.LLM	0.5 (48%)	—	—
ThinkPrune-1.5B / MATH500	SnapKV	—	—	—
	TOVA	—	—	—
	Str.LLM	—	—	—
DeepSeek-7B / MATH500	SnapKV	0.5 (49%)	0.25 (24%)	—
	TOVA	0.5 (49%)	—	—
	Str.LLM	0.5 (49%)	—	—
DeepSeek-7B / MMLU-Pro	SnapKV	0.25 (24%)	—	—
	TOVA	0.25 (24%)	—	—
	Str.LLM	—	—	—

AB SERVING MICROBENCHMARK: BATCHED DECODING FROM COMPRESSED CACHES

Mirroring a serving stack, each request's midpoint context from the registered Llama-8B preflight-valid pool is prefilled and compressed individually, the caches are padded into one batched cache, and the fixed elicitation is decoded greedily for the whole batch (64 new tokens; single L40S; four

batches per cell). Retention is unconditional answer correctness on the sampled rows (not the conditional survival S of Eq. 2); only comparisons at the same batch size are interpretable, because different batch sizes sample different prefixes. Because batches index the same rows across policies, every compressed batch has a paired no-compression comparator at the same batch size and index, so we report interval estimates rather than win counts (`src/analyze_serving_stats.py`, `paper/serving_stats_results.json`). Pooling the twelve matched batches, the calibrated 90%-retention policy (SnapKV r=0.25) raises decode throughput by 23.6% (stratified bootstrap 95% CI 21.9–25.2, resampling batches within batch size, 10^5 draws) and lowers peak decode memory by 8.1% (7.3–8.9), winning 12 of 12 matched batches on each (two-sided sign $p = 4.9 \times 10^{-4}$); both effects grow with batch size ($+18.0$/$+24.5$/$+28.5$% throughput and -4.0/-8.3/-11.8% peak memory at batches 8/16/24, 4 of 4 wins in every cell). Uniform r=0.5 is faster still ($+61.4$%, CI 56.5–66.1) and uncertified. *We make no accuracy comparison here*: each cell scores 32 to 96 rows of unconditional correctness (calibrated versus none: 17/32, 40/64, 61/96 against 19/32, 41/64, 60/96), which supports neither a difference nor an equivalence claim. The memory gain restated as capacity: a ∼30 GB budget fits batch 16 uncompressed versus batch 24 calibrated. Full per-cell measurements: Table 38; machine-readable report: `paper/serving_bench_results.json`.

Table 38: Batched answer decoding from individually prefilled caches (Llama-8B, GSM8K midpoints; four batches per cell). Cache, tokens/s and accuracy are means (accuracy pooled over the batches' rows); "Peak" is the maximum peak decode memory over the four batches.

Policy	Batch	Cache (GB)	Peak (GB)	Tokens/s	Accuracy
no compression	8	2.18	22.2	142	59.4%
no compression	16	6.27	27.7	133	64.1%
no compression	24	11.55	34.2	109	62.5%
uniform SnapKV r=0.5	8	1.09	19.2	200	50.0%
uniform SnapKV r=0.5	16	3.13	22.0	214	57.8%
uniform SnapKV r=0.5	24	5.78	25.3	195	60.4%
calibrated SnapKV r=0.25	8	1.63	20.8	166	53.1%
calibrated SnapKV r=0.25	16	4.70	24.9	164	62.5%
calibrated SnapKV r=0.25	24	8.66	29.8	139	63.5%

AC FEW-SHOT BUDGET CALIBRATION VIA EMPIRICAL-BAYES SHRINKAGE

The calibrated budgets of Section 4.5 need on the order of a hundred calibration questions per (model, task) pair. This exploratory analysis tests whether the registered grids can serve as a prior so that *another model–task grid within the pooled benchmark families* needs only a handful: leave-one-grid-out over the thirteen grids, a moment-matched Beta prior per (compressor, ratio) fitted on the remaining grids (strength capped at 50 pseudo-counts), updated by a k-shot calibration sample (200 deterministic replicates), with the fixed-sequence walk run on posterior δ=0.05 quantiles. The posterior quantile is a Bayesian credible bound, *not* a frequentist guarantee, so the shrinkage rule *selects* rather than certifies; the comparator runs small-sample Clopper–Pearson (which does certify) on the identical shots. Table 39 reports the trade-off: at 20 shots the shrinkage rule selects nonzero compression in 55.4% of the 78 held-out cells against 28.7% for the naive rule, and at 10 shots the naive rule can certify *nothing* while shrinkage still selects 45.1%, at 98.6% mean held-out coverage. The cost is concentrated where the prior misleads (MMLU-Pro's reversed ranking, DeepScaleR): worst-case per-cell coverage falls to 77.5%, with 10 of 78 cells below 0.95.

Two sensitivity analyses bound the scope of this transfer. Excluding the held-out *task* from the prior pool leaves the average utility–coverage trade-off nearly unchanged (mean coverage 97.6%, worst case 77.5%, nonzero selections 55.5%) but does not restore per-cell coverage. Excluding the held-out *model family* leaves, for the twelve Qwen-family grids, a prior pool of a single Llama grid, and the result collapses (mean coverage 85.6%, worst case 2.5%, 26 of 78 cells below 95%, with inflated 82.4% nonzero selections): transfer from the single available cross-family prior is unreliable, and broader family coverage would be needed before any cross-family claim. The pooling is within-benchmark/within-family, not a claim about arbitrary new pairs. The prior-strength cap matters at the tail: caps of 5–20 pseudo-counts keep worst-case coverage at 88.0–88.5% with nearly identical selections (52.8–54.6%), so the reported cap of 50 is the most aggressive, not the safest, setting (machine-readable report: `paper/fewshot_budget_results.json`).

Table 39: Leave-one-grid-out few-shot calibration (δ=0.05, targets 90/75%, 200 replicates, 78 cells). Coverage = share of replicates whose deployed budget meets the target on the remaining eligible questions; nonzero = share selecting any compression; match = agreement with the full-sample budget.

Shots	Rule	mean cov.	min cov.	nonzero	match
10	shrinkage	0.986	0.720	0.451	0.480
	naive CP	1.000	1.000	0.000	0.269
20	shrinkage	0.980	0.775	0.554	0.628
	naive CP	0.996	0.955	0.287	0.384
40	shrinkage	0.974	0.670	0.620	0.741
	naive CP	0.986	0.900	0.481	0.601

AD LENGTH-QUARTILE AND PER-TRAJECTORY DISTRIBUTION FIGURES

Figures 9 and 10 accompany Section 4.4: the length-quartile view of the mean critical ratio, and the per-trajectory critical-ratio distributions that the aggregate hides. Within each grid's eligible set, no positive Spearman association between midpoint length and critical ratio survives the 39-cell Bonferroni threshold, and the only two survivors (Qwen3-8B/TOVA, L1-Qwen3-8B/TOVA, $\rho \approx -0.21$) are *negative*. The quartile view agrees: token spans widen up to 25× while mean critical ratios move by at most 0.12. The per-trajectory distributions are nevertheless wide, sometimes bimodal: DeepSeek-7B keeps 49% of eligible SnapKV trajectories at r^*=0.9 and 11% at $r^* \leq 0.25$, exploitable variance that the method gate of Section 4.4 fails to convert.

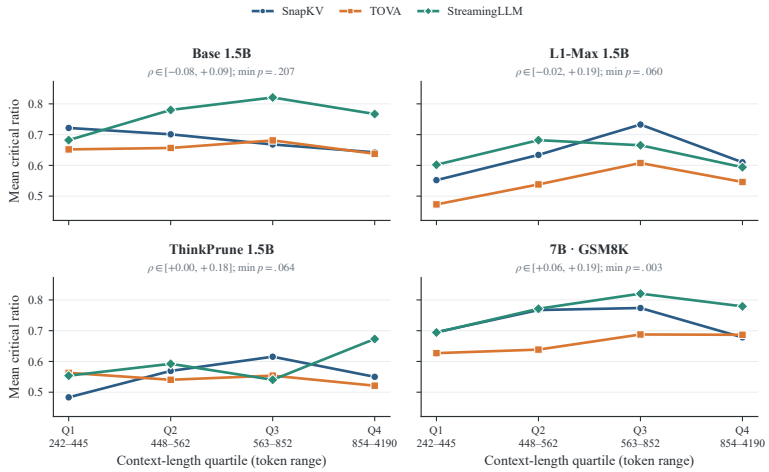


Figure 9: Mean critical compression ratio by formatted-length quartile (Q1–Q4) for the four GSM8K grids. Token spans widen by up to 25× from Q1 to Q4 (annotated), yet mean critical ratios move by less than 0.12 in every model–compressor cell, with no consistent direction. Panel titles summarize the compressor-specific ranges of Spearman ρ and permutation p for the corresponding grid.

AE MATCHED-SUBSET SURVIVAL ON GSM8K

Per-model conditional survival conditions on each model's own eligible set, and those sets differ in size and composition (212/102/105 on GSM8K for the 1.5B trio), so cross-model survival statements could in principle reflect selection rather than robustness. To control for this, we match questions by text across the four originally registered GSM8K grids and restrict to the n=54 questions that are jointly source-correct and jointly ratio-zero-eligible for every model, so all models are compared on the identical question set. Table 40 reports survival on this joint subset with paired McNemar exact tests against Base (nominal, uncorrected): the Base advantage over both length-penalty checkpoints persists in most cells (e.g., at 75% StreamingLLM compression, Base retains 96.3% versus 72.2% for L1-Max, $p = 0.001$, and 59.3% for ThinkPrune, $p < 10^{-4}$), while Base and DeepSeek-7B never differ significantly. These matched results are inconsistent with a selection explanation of the survival gap, but with n=54 and nominal per-cell tests they remain supporting rather than definitive evidence (machine-readable report: `paper/joint_survival_results.json`).

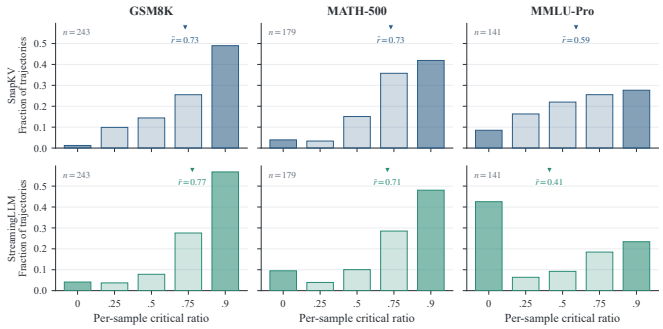


Figure 10: Distribution of per-example critical compression ratios over each grid's eligible set (SnapKV top row, StreamingLLM bottom row; DeepSeek-7B on GSM8K, MATH-500, and MMLU-Pro). Aggregate mean critical ratios hide substantial per-trajectory structure: large mass concentrates at both extremes, and the MMLU-Pro StreamingLLM cell is visibly bimodal.

Table 40: Conditional survival on the 54 jointly source-correct, jointly eligible GSM8K questions. *: paired McNemar exact $p < 0.05$ versus Base in the same cell.

	ratio 0.75			ratio 0.9		
Model	SnapKV	TOVA	Str.LLM	SnapKV	TOVA	Str.LLM
Base	85.2%	74.1%	96.3%	66.7%	53.7%	64.8%
DeepSeek-7B	79.6%	66.7%	90.7%	51.9%	50.0%	59.3%
L1-Max	70.4%	46.3%*	72.2%*	42.6%*	29.6%*	46.3%
ThinkPrune	46.3%*	55.6%	59.3%*	11.1%*	31.5%*	40.7%*

AF LONG-TRACE PILOT: COMPETITION PROBLEMS AT 8–16K REASONING TOKENS (EXPLORATORY)

An exploratory pilot extends the protocol to long traces (specification frozen before any run): the registered Llama-8B base generates completed trajectories for AIME 2024/2025 and AMC 2023 (100 questions; question file pinned by hash) at a 14,336-token budget with the registered sampling configuration, and the audited midpoint pipeline runs unchanged (98 pass preflight; 51 source-correct; full compressor grid; report `paper/long_pilot_results.json`). Median reasoning length is 10,480 tokens (versus 1,016 on the same model's registered GSM8K grid) and median midpoint KV 2,004 tokens (versus 563).

Three observations, all on small samples (51 source-correct rows; 16 eligible for survival; Wilson intervals quoted; Table 41). (i) Midpoint recoverability is far lower on these long, hard problems for the *same* model: $R = 31.4\%$ [20.3, 45.0] versus 74.3% [68.8, 79.1] on GSM8K, consistent with Section 4.3: extractable answers concentrate late, and on competition problems the 50% state rarely suffices (task difficulty and trace length are entangled here; the pilot does not separate them). (ii) The compressor ranking shifts again: TOVA is strongest (mean critical ratio 0.738), StreamingLLM intermediate (0.659), SnapKV weakest (0.600), against StreamingLLM-first on the same model's GSM8K grid, reinforcing that compressor conclusions do not transfer across task distributions (Section 4.2). (iii) The survival structure itself persists at this scale: graded decay with nonzero critical ratios under every compressor. No cross-model, causal, or calibration claim is made at this scale; the pilot bounds none of the registered claims and motivates a full long-context study as future work.

Table 41: Long-trace pilot versus the same model's registered GSM8K grid (exploratory; recoverability on source-correct rows, survival on the n=16 eligible rows).

Quantity	GSM8K (registered)	AIME/AMC pilot
questions (source-correct)	300 (276)	98 (51)
$R(0.5)$ at ratio zero	74.3% [68.8, 79.1]	31.4% [20.3, 45.0]
median reasoning tokens	1,016	10,480
median midpoint KV tokens	563	2,004
mean critical ratio (SnapKV/TOVA/Str.LLM)	0.69/0.74/0.78	0.60/0.74/0.66
$S(0.75)$ (SnapKV/TOVA/Str.LLM)	76/76/86%	50/75/75%

AG A WORKED MIDPOINT EXAMPLE

On a GSM8K house-flipping question ($80,000 purchase, $50,000 repairs, value increased by 150%; gold profit $70,000), the Base midpoint (2185 retained tokens) elicits the correct profit, while the ThinkPrune midpoint (298 retained tokens) commits to $200,000, the increased house value: an intermediate quantity already written in the retained prefix, reported in place of the requested profit before the final subtraction is carried out. The prefix writes that quantity as "$200,000"; the separator-tolerant rule v2 of Appendix L matches that comma-grouped form, so this item *is* one of the in-prefix errors counted below. (Under the superseded rule v1, which matched only the bare digit string, it was filed under *novel*; it is one of the 8 rows the rule-v2 upgrade moved.) It is a single item chosen for readability, and it should not be read as the general mechanism behind the RL deficit. Applying the same string rule to *every* wrong uncompressed elicitation (Appendix L) shows that closing on a number already present in the retained prefix is still the *minority* of RL midpoint errors, 17.6% for L1-Max and 25.6% for ThinkPrune on GSM8K, against 39.3% for Base, while the majority (68.8% and 63.5%) are numbers appearing nowhere in the model's own trajectory. The example illustrates one real category of failure, not the modal one.

AH ENGAGED DECODE-TIME BENCHMARK

An earlier decoding-only benchmark evicted cache entries only after the boxed answer had been emitted and is withdrawn in the invalidation ledger. The corrected protocol compresses the midpoint context *during prefill* with the same base compressor and continues eviction every 32 decode steps; every record stores the first-eviction phase and step and the token index at which the boxed answer closes. Across all 1800 compressed runs (300 GSM8K Base midpoint contexts, three presses, ratios 0.5 and 0.75, 256-token elicitations), the eviction-precedes-answer assertion holds in 1800/1800.

Table 42: Engaged decode-time benchmark, prefill-decoding protocol (300 contexts per cell; label flips count records discordant with the ratio-0 run of the same press, in either direction; agreement pairs each cell with the fixed-prefill grid).

Press	ratio	correct/300	label flips	prefill agreement
StreamingLLM	0	212	—	—
StreamingLLM	0.5	198	20	300/300
StreamingLLM	0.75	184	32	300/300
TOVA	0	212	—	—
TOVA	0.5	183	39	300/300
TOVA	0.75	137	85	300/300
Knorm	0	212	—	—
Knorm	0.5	47	171	—
Knorm	0.75	15	201	—

Under real pre-answer compression, damage appears immediately (Table 42): StreamingLLM flips 20/32 correctness labels at 50%/75% compression, of which 17/30 are losses, and TOVA 39/85 of which 34/80 are losses; and Knorm, a compressor with no prefill-grid counterpart, collapses to 47/300 and 15/300. For the two presses shared with the fixed-prefill grid, per-cell outcomes agree in all 900 paired cells per press (1,800/1,800 overall): under a deterministic greedy stack, the fixed-prefill protocol is exactly representative of an engaged decoding stack whose cache is compressed before answer emission. Peak memory (7.96 GB) and throughput (60–62 tokens/s) are flat across ratios at this scale, so this is correctness evidence, not a serving win. Because the boxed answer closes within a few tokens, repeated long-horizon decode-time eviction is exercised only after the answer; that regime remains untested.

RL checkpoints under the identical engaged protocol. We reran the protocol unchanged (same seed, ratios, 32-step eviction interval, pinned revisions) on L1-Max and ThinkPrune, 300 midpoint contexts each (Table 43); the eviction-precedes-answer assertion holds in all 3,600 compressed RL runs, and 4 of 8,100 records across the nine cells are uncommitted (counted incorrect). Three findings. (i) At ratio zero, meaning a free 256-token continuation from the re-encoded midpoint with no compression and no forced prefill beyond the shared elicitation, the decode-time construction reproduces the recoverability gap: 212/300 for Base versus 108 and 105. (ii) Under engaged eviction the

gap is not an artifact of the 64-token elicitation: among jointly ratio-zero-correct pairs, Base's continuation survives eviction better in all eight StreamingLLM/TOVA contrasts (91.3–97.8% vs. 58.7–77.8% at r=0.5–0.75 for StreamingLLM; paired McNemar $p \leq 0.017$ in every cell). (iii) Knorm floors all three models (conditional survival 6–21%); at r=0.75 the Base-versus-RL contrast is no longer detectable there ($p \geq 0.26$), a floor effect rather than RL robustness. RL runs executed on a consumer RTX 5060 Ti and an L40S (BF16; torch 2.9.1/2.5.1); machine-readable report: `paper/decode_rl_results.json`.

Table 43: Engaged decode-time eviction, RL checkpoints versus Base (300 contexts per cell; correct counts, and conditional survival among each cell's ratio-zero-correct rows).

Model	Press	ratio 0	surv. 0.5	surv. 0.75
Base	StreamingLLM	212/300	92.0%	85.8%
Base	TOVA	212/300	84.0%	62.3%
Base	Knorm	212/300	20.8%	6.1%
L1-Max	StreamingLLM	108/300	75.0%	63.9%
L1-Max	TOVA	108/300	67.6%	45.4%
L1-Max	Knorm	108/300	11.1%	6.5%
ThinkPrune	StreamingLLM	105/300	77.1%	60.0%
ThinkPrune	TOVA	105/300	67.6%	49.5%
ThinkPrune	Knorm	105/300	11.4%	8.6%