PROCEDURE OR FORMAT? MEASURING WHAT A REASONING TRACE TEACHES WITH FORMAT-RESIDUALIZED CODELENGTH

Anonymous authors
Paper under double-blind review

ABSTRACT

A chain-of-thought trace can lower a model's loss on related problems for two different reasons: it can teach a procedure that transfers, or it can merely prime the surface format those problems share. A score based on that loss reduction cannot tell the two apart, and we prove that no function of it can. We therefore measure the reduction twice, once on sibling problems that need the trace's procedure and once on a matched null set in the same format that needs a different one, and take the difference. Format priming is shared by both sets and cancels; we call what remains the format-residualized differential codelength R. Under the standard mixture view of in-context learning, R is, up to a separation term, the expected log Bayes factor the trace induces between its own procedure and the null's. On 224 clusters built from GSM-Symbolic, R separates valid from wrong-method reasoning by 0.251 bits per token ($p = 10^{-5}$) with trace length and answer shape controlled, and where the raw reduction ranks a fluent trace that performs no computation above a valid one, R ranks them correctly ($p = 9 \times 10^{-7}$). On 216 controlled clusters in three domains under five open models, R falls with corruption severity and predicts an independent model's out-of-distribution transfer beyond length, raw reduction, and correctness (partial $\rho = 0.27$, $p = 2 \times 10^{-15}$). The result is a label-free, forward-pass-only test of whether a trace teaches the procedure a problem family needs. It requires a scorer that has not already solved the task, and it measures in-context value, which does not carry over to selecting fine-tuning data. Code and data are available at https://anonymous.4open.science/r/format-residualized-codelength-8D79.

1 INTRODUCTION

Chain-of-thought prompting improves the problem-solving ability of large language models (Wei et al., 2022; Kojima et al., 2022; Nye et al., 2021), but its value is uneven across tasks (Sprague et al., 2025), and not every trace that lowers a model's loss teaches something transferable. A large literature therefore scores individual reasoning steps so that good reasoning can be rewarded, selected, or distilled: by plausibility, with process reward models trained on human or automatic step labels (Lightman et al., 2024; Wang et al., 2024; Uesato et al., 2022), or by outcome, crediting every step on a trace that reaches the correct answer (Cobbe et al., 2021). Both are confounded. Plausibility scores are vulnerable to reward hacking (Gao et al., 2023) and generalize unevenly (Zheng et al., 2025), and outcome credit rewards luck along a correct path while penalizing useful reasoning on an incorrect one. Neither answers the question that actually governs generalization, how much a trace teaches a transferable procedure as opposed to priming a surface format, so optimizing, ranking, or distilling on either signal can propagate recitation as readily as reasoning.

The task we address is to assign each reasoning trace a continuous coordinate for how much it helps a model solve structurally related but distinct problems. The confound that makes this nontrivial, a trace lowering loss either by teaching the transferable principle the family requires or by merely priming the family's surface format, is the same one that drives the fragility documented by symbolic and counterfactual perturbation studies, where accuracy collapses under changes that leave the underlying procedure intact (Mirzadeh et al., 2025; Wu et al., 2024), and it is closely related to evidence that chain-of-thought explanations do not always reflect the computation that produced an

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.

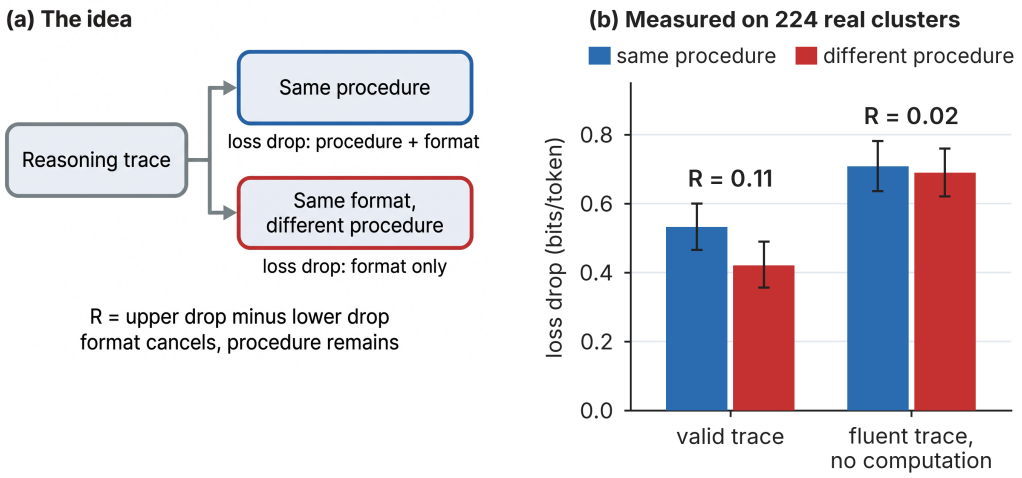

Figure 1: (a) Schematic: a trace is prepended to two problem sets sharing a surface format; format priming lowers the answer codelength of both, the procedure only that of the set needing it, and R keeps the difference. (b) Measured on all 224 real GSM-Symbolic clusters under Llama-3.1-8B (Section 6.2), means with 95% bootstrap intervals: a fluent trace that performs no computation earns the *larger* reduction on the same-procedure set, so the raw reduction prefers it, but it helps the null set almost equally and its R is near zero. R ranks the valid trace above it in 63% of clusters; the raw reduction does so in 24%.

answer (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025). Separating these two at the level of an individual trace is exactly what existing reward, verification, and influence signals do not do, and what a measure of method-value must.

Our starting point is the equivalence between in-context learning and prequential, or online, coding: presenting examples in context and measuring the model's compression of subsequent answers is a description-length measurement (Xie et al., 2022; Delétang et al., 2024; Blier & Ollivier, 2018), so a trace that genuinely teaches a procedure should compress the answers of problems that need it. The methodological contribution of this paper is to turn the format confound from a nuisance into the measurement. We score a trace by the teacher-forced answer-codelength reduction it induces on a set of procedure-shifted siblings, then subtract the reduction it induces on a format-matched, difficulty-matched null set whose required procedure differs. Because both sets share surface format and receive the identical prepended trace, any format priming lowers the codelength of both equally and cancels in the difference (Figure 1a). What survives, we argue and then test, is the transferable method-value of the trace: its content about the procedure the family requires (Section 3).

We make five contributions. (i) We define R and establish its theory: it is the log Bayes factor a trace induces between the demonstrated and the null procedure, which turns format cancellation from an assumption into a corollary and makes the capacity condition quantitative; the residual is provably necessary, since no statistic of the raw reduction alone identifies method-value; and we test the theory's own idealization rather than assuming it (Sections 2, 3). (ii) Across 216 controlled principle-clusters in three domains under five open models, and across 224 clusters built entirely from the public GSM-Symbolic benchmark, R is graded with corruption severity and separates valid from wrong-method reasoning (Sections 6.3, 6.1). (iii) Using an independent transfer model never used for scoring, R predicts realized out-of-distribution transfer beyond trace length, raw un-residualized codelength reduction, correctness, and a process reward model it complements, the two measuring different constructs (Sections 6.4, 6.5). (iv) We isolate what is regime-dependent: under a compositional rather than faithful shift of the principle the marginal relation between codelength and transfer reverses, but the reversal is carried by the raw un-residualized reduction, and the residual's incremental contribution remains positive in both regimes (Section 6.4). (v) We characterize the method's scope honestly: a useful scorer must not already solve the task (Section 6.3), and R measures in-context value, which we show does not transfer to selecting fine-tuning data (Section 6.6). Related work follows in Section 4, once the construct is on the table.

2 METHOD

Codelength. Let M be a scorer model, x a problem with gold answer tokenized as $a = (a_1, \ldots, a_{|a|})$, and c an in-context prefix. The per-token teacher-forced answer codelength is

$$L_M(a \mid x, c) \;=\; -\frac{1}{|a|} \sum_{k=1}^{|a|} \log_2 p_M(a_k \mid c, x, a_{<k}) \,, \tag{1}$$

the average number of bits to encode the answer given the context. It is obtained from a single forward pass over the concatenation of c, x, and a, with no autoregressive generation. For a set $\mathcal{S}$ of problems we write $L_M(\mathcal{S} \mid c)$ for the token-weighted mean of (1) over its members.

Principle-clusters and the two sibling sets. A principle-cluster is a single transferable procedure instantiated in a fixed surface frame. Within a frame, the question selects the procedure, so two sets share format while differing in the required procedure. The procedure-shifted set $\mathcal{P}$ contains problems for which the trace's principle applies; the null set $\mathcal{N}$ contains problems in the same frame whose required procedure differs, matched to $\mathcal{P}$ in answer-token length and baseline difficulty; we call the members of $\mathcal{P}$ and $\mathcal{N}$ the cluster's siblings. A candidate trace t is a worked solution to a seed problem of the cluster.

Format-residualized differential codelength. Let c_0 be a fixed baseline context, identical for $\mathcal{P}$ and $\mathcal{N}$. The drop a trace induces on a set is $\Delta L(\mathcal{S} \mid t) = L_M(\mathcal{S} \mid c_0) - L_M(\mathcal{S} \mid t)$, and the residual method-value is the difference in drops,

$$R_M(t) \;=\; \Delta L(\mathcal{P} \mid t) - \Delta L(\mathcal{N} \mid t) \;=\; \big[L_M(\mathcal{P} \mid c_0) - L_M(\mathcal{N} \mid c_0)\big] - \big[L_M(\mathcal{P} \mid t) - L_M(\mathcal{N} \mid t)\big]. \tag{2}$$

The residual construction is designed to remove a format-priming term. Earlier work of ours assumed an additive decomposition to justify this; Section 3 instead *derives* it, as Corollary 1 of a characterization of what R measures, so cancellation is a consequence of a standard model of in-context learning rather than a postulate. Note that difficulty matching does not make the baseline gap negligible in practice (its mean magnitude is 0.33 to 0.62 bits per token across scorers); what protects the measurement is not a small gap but the fact that the gap is trace-independent, so it cancels exactly in any contrast between traces (Proposition 2).

Remark 1. *The negative-knowledge contrast, $R_M(t^+) - R_M(t^-)$ for a valid trace t^+ and a fluent but wrong trace t^- within the same cluster, removes the baseline term entirely, since both traces are scored against the same $\mathcal{P}$, $\mathcal{N}$, and c_0. It is therefore robust to residual difficulty imbalance.*

Three controls. We rely on three controls: the residual in (2); a second, independent scorer family, since a genuine signal is a property of the trace and agrees across families while a scorer-circular artifact does not; and a negative-knowledge control, a fluent wrong-method trace that must score below a valid one in the same cluster. A per-step overlay, the prequential view of (2), follows from scoring R on trace prefixes (Appendix C).

3 THEORETICAL ACCOUNT

Having defined R, we now characterize what it measures and when it can be trusted. The central result is mechanistic: under the standard Bayesian-mixture view of in-context learning, R is the log Bayes factor a trace induces between the demonstrated procedure and the null procedure. Format cancellation and the capacity condition then follow as corollaries rather than as assumptions, and an identifiability result shows the residual is necessary and not merely helpful. Three supporting results, including an assumption-free divergence reading and a concentration rate, are summarized at the end of the section and developed in the appendix. Throughout, $\mathcal{D}_z(\cdot \mid x)$ is the answer law of a problem x under procedure z, and for a context c the per-example codelength is the cross-entropy $H(\mathcal{D}_z(\cdot \mid x),\, p_M(\cdot \mid x, c))$. We write

$$\delta_t(x) = H\big(\mathcal{D}(\cdot \mid x),\, p_M(\cdot \mid x, c_0)\big) - H\big(\mathcal{D}(\cdot \mid x),\, p_M(\cdot \mid x, t)\big) \tag{3}$$

for the per-problem codelength reduction a trace induces, so $\Delta L_M(\mathcal{S} \mid t) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \delta_t(x)$, the empirical $R_M(t) = \Delta L_M(\mathcal{P} \mid t) - \Delta L_M(\mathcal{N} \mid t)$ of (2), and the population value $R^*_M(t) = \mathbb{E}_{z^*}[\delta_t] - \mathbb{E}_{z_N}[\delta_t]$, where $\mathcal{P}$ and $\mathcal{N}$ are n problems drawn independently from the trace's procedure z^* and a different procedure z_N. We define the *method-value* of t as the information it carries about z^*, the transferable procedure its cluster requires; R is its format-residualized estimator, the sense in which R measures method rather than execution throughout. All proofs are in Appendix B.

R is a log Bayes factor over procedures. The divergence view says what R measures but not why the residual isolates method. For that we use the Bayesian-mixture account of in-context learning this paper already builds on (Xie et al., 2022; Saunshi et al., 2021), under which $p_M(a \mid x, c) = \sum_z p(a \mid x, z)\, \pi(z \mid c, x)$ for a posterior π over procedures. Say the procedures in a cluster are ε-separated if, for an answer generated by z, the mass the other procedures place on it is at most an ε fraction of the true one, under both contexts. This is what the benchmark's construction enforces, since $\mathcal{P}$ and $\mathcal{N}$ differ in the required procedure.

Theorem 1 (Bayes-factor characterization)**.** *Under ε-separation,*

$$R^*_M(t) \;=\; \mathbb{E}_x \log_2 \frac{\pi(z^* \mid t, x)}{\pi(z^* \mid c_0, x)} \;-\; \mathbb{E}_x \log_2 \frac{\pi(z_N \mid t, x)}{\pi(z_N \mid c_0, x)} \;+\; O(\varepsilon),$$

the expected log Bayes factor the trace induces between the demonstrated and the null procedure.

The proof splits the mixture at the generating procedure, at which point the instance likelihood $p(a \mid x, z)$ appears in both contexts and cancels *exactly*. That cancellation is the formal content: to the extent the mixture model holds, R does not report how the answer is distributed under the procedure, only how the trace moved the model's belief about which procedure applies. Two consequences replace assumptions made elsewhere in the literature and in earlier versions of this work.

Corollary 1 (Format cancellation)**.** *If a trace is* format-only *for the frame, meaning it reweights every procedure in that frame by a common factor, then the two ratios coincide and $R^*_M(t) = O(\varepsilon)$.*

This gives "format priming" a precise meaning, a trace carrying no information about *which* procedure the frame requires, and derives the cancellation Section 2 would otherwise assume.

Corollary 2 (Capacity)**.** *Write $s_0 = -\log_2 \pi(z^* \mid c_0)$ for the scorer's baseline surprisal about the correct procedure. If the trace does not suppress the null procedure below a fraction γ of its baseline mass, then $R^*_M(t) \le s_0 + \log_2(1/\gamma)$.*

Achievable separation is thus capped by how much posterior mass the scorer has *not yet* placed on the right procedure: a scorer that already solves the task has $s_0 \to 0$ and cannot separate traces. This is sharper than noting that a reduction cannot exceed its baseline, and Section 6.3 tests it.

The residual is necessary, not merely helpful.

Theorem 2 (Identifiability)**.** *Under the same model there exist traces t_1, t_2 with identical raw reduction $\Delta L_M(\mathcal{P} \mid t_1) = \Delta L_M(\mathcal{P} \mid t_2)$ but $R^*_M(t_1) = 0 < R^*_M(t_2)$. Hence no function of $\Delta L_M(\mathcal{P} \mid \cdot)$ alone can separate format priming from method-value, while R does.*

The witness is explicit: a trace that raises every in-frame procedure's posterior by a common factor produces the same raw reduction as one that raises only $\pi(z^*)$ by that factor, and the two are distinguished only by the contrast. This turns the empirical observation that R predicts transfer beyond the raw reduction (Section 6.4) into a confirmation of an identifiability result rather than a fortunate regression.

Theorem 1 is stated in expectation over x deliberately: the stronger claim that the log-ratio is constant across a cluster's siblings is testable and we find it false, though the exact cancellation of $p(a \mid x, z)$ survives and is all the expectation form requires (Appendix C.3).

Three supporting results. Three further results are used but not developed here. R admits an assumption-free reading as a differential reduction in divergence from the true answer law (Proposition 1). Its cluster term is trace-independent, so contrasts between traces are exactly baseline-invariant, which is what licenses the cluster-centered statistic we use when pooling across clusters (Proposition 2). And R concentrates at $O(1/\sqrt{n})$ in the number of siblings, a rate we verify, while

an exchangeability identity bounds faithful-transfer codelength reduction below by R on scorers that compress rather than lengthen the null set (Theorem 3, Theorem 4). Statements, proofs, and the compositional remark are in Appendices A and B.

4 Related Work

Existing ways of scoring a trace measure something correlated with its value but not the thing itself, each conflating procedure with presentation differently. We group them by what they measure and state what remains unmeasured.

Signals that score correctness. Process reward models judge step correctness from human or search-derived labels (Lightman et al., 2024; Wang et al., 2024; Uesato et al., 2022; Zhang et al., 2025b), increasingly as generative critics (Zhang et al., 2025a); outcome supervision credits every step on a trace that reaches the right answer (Cobbe et al., 2021); and both now drive test-time scaling and reasoning training (Snell et al., 2025; Guo et al., 2025; Yao et al., 2023), though step-correctness models generalize unevenly across difficulty (Zheng et al., 2025). Closest to our label-free setting, implicit process rewards read a per-step signal from the log-likelihood ratio of a response-level reward model (Yuan et al., 2025; Cui et al., 2026). All target whether a trace is *right*, which is not whether it is *instructive*. The two come apart measurably: a right-method, wrong-number trace is scored lowest by a process reward model yet transfers as well as a valid one, so correctness-based curation discards what it should keep (Section 6.5).

Signals that score compression. Reading a model's compression of later answers after a trace is a prequential, minimum-description-length measurement (Dawid, 1984; Rissanen, 1978; Grünwald, 2007), the view behind language-modeling-as-compression (Delétang et al., 2024; Blier & Ollivier, 2018), the compression-capability tie (Huang et al., 2024), and description-length probes of representations and datasets (Voita & Titov, 2020; Perez et al., 2021); predictive $\mathcal{V}$-information is the closest relative (Xu et al., 2020; Ethayarajh et al., 2022). Compression is the right currency and we adopt it, but an un-residualized reduction cannot say *why* it fell. This is not fastidious: on real traces the raw reduction ranks a fluent trace performing no computation *above* a valid one, because fluent prose in the family's format is itself compressive (Section 6.2). Theorem 2 shows this is an identifiability failure, not a tuning one.

Signals that score influence, and what the confound costs. Value-of-an-example methods attribute behaviour to training points via influence functions, gradient tracing, or Shapley valuation (Koh & Liang, 2017; Pruthi et al., 2020; Ghorbani & Zou, 2019), select demonstrations by similarity or utility (Liu et al., 2022), and distil rationales into weights (Hsieh et al., 2023; Zelikman et al., 2022). All need gradients, retraining, or labels, which puts them out of reach for scoring one trace at curation time. Meanwhile the confound is well documented, with accuracy collapsing under variants that leave the procedure intact (Wu et al., 2024; Mirzadeh et al., 2025; Dziri et al., 2023) and stated chains of thought not always reflecting the computation behind the answer (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025). The field thus has strong evidence that models lean on non-transferable surface procedures, and no per-trace instrument separating the two. R is that instrument: it needs one forward pass and no labels, it is format-residualized by construction rather than by correction, and the latent-concept view of in-context learning (Xie et al., 2022; Min et al., 2022; Olsson et al., 2022; von Oswald et al., 2023) is what lets us say precisely what it measures, namely the log Bayes factor between the demonstrated procedure and a null one (Theorem 1).

5 Experimental Setup

Benchmark. We construct 216 principle-clusters as three random seeds crossed with six instances per principle, four principles, and three domains: arithmetic word problems, relational kinship deduction, and algorithmic list tracing. Each cluster provides a seed problem, a valid trace and a five-level corruption ladder (an arithmetic-error variant that keeps the method but changes the result, a wrong-operation variant, a fluent non-sequitur, a sentence-shuffled variant, and a truncated variant), ten procedure-shifted siblings, ten null siblings matched by answer-digit count and morphology,

twelve faithful out-of-distribution (OOD) problems that shift the principle to a harder instance, and twelve compositional OOD problems that compose the principle with a new operation.

Scorers and models. We score codelength in 32-bit precision under five open models spanning three families and four sizes: Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Llama-3.1-8B, and Mistral-7B (Qwen Team, 2024; Grattafiori et al., 2024; Jiang et al., 2023), computed with vLLM prompt log-probabilities (Kwon et al., 2023) as in Algorithm 1. Realized transfer is measured on Mistral-7B, with Llama-3.1-8B as a second in-band model (Appendix C); when predicting transfer we standardize R within each scorer and average over the scorers excluding the transfer model, so neither a scorer's own transfer nor its scale drives the average. On the real clusters the valid trace is the benchmark's own worked solution and the arithmetic-error variant is derived from it programmatically; the wrong-method and non-sequitur traces and the rewritten null questions come from Qwen2.5-7B-Instruct, and the pool of Section 6.6 adds Qwen3-8B and SmolLM3-3B. We compare against Qwen2.5-Math-PRM-7B (Zhang et al., 2025b) as a process reward baseline on the arithmetic domain. We validate the measurement itself at scale rather than on isolated examples: reconstructing the set-level baseline codelengths from independently recomputed per-problem values under a Hugging Face forward pass reproduces the vLLM values to a mean of 0.0006 and a maximum of 0.0051 bits across 144 cluster-sets, while 16-bit kernels disagree by up to 0.17 bits. We additionally disable TensorFloat-32 matrix multiplication, which some builds enable by default beneath a nominal 32-bit dtype and which by itself perturbs answer codelength by up to 0.006 bits, comparable to the entire engine tolerance; all scoring therefore uses true 32-bit precision.

Transfer measurement. The transfer gain of a trace on a held-out problem is the transfer model's accuracy with the trace as a worked example minus its accuracy with no example; the transfer model is always distinct from the scorer, breaking scorer-circularity.

6 Results

We establish the separation on real data, isolate what the residual adds over raw compression, use the templated benchmark for graded manipulation, show R predicts realized transfer, and close with its boundaries.

6.1 Two real benchmarks, two scorer families

We begin on data we did not construct. In GSM-Symbolic (Mirzadeh et al., 2025), whose templates instantiate one GSM8K-derived procedure in a fixed frame with varying names and numbers, a template *is* a principle-cluster: $\mathcal{P}$ is other instances of the same template and the trace is the benchmark's own worked solution. Both OOD regimes also come from the published benchmark, since held-out instances shift the instance while the `p1` configuration adds a clause, composing the principle with a new operation.

No single null is clean on real data, so we use two whose biases run in opposite directions and require the conclusion to hold under both (Appendix C.2).

Across 224 real clusters under Llama-3.1-8B, R separates valid from wrong-method reasoning by 0.474 bits per token ($[0.383, 0.568]$, $p = 5 \times 10^{-18}$) under the conservative null and 0.588 ($[0.501, 0.678]$, $p = 2 \times 10^{-26}$) under the cross-template null, ordering as designed. The clusters are template-by-seed pairs over 79 distinct templates, so we also collapse to the template before testing; the separation is unchanged ($+0.469$, $p = 3 \times 10^{-10}$). That figure is inflated by a confound we then remove. Because only the gold answer is scored, a trace whose final number has a different digit count from the sibling answers is penalized for its answer's *shape* rather than its method, and wrong-method traces mismatch far more often than valid ones (mean absolute digit gap 0.822 against 0.261). Restricting to the 92 clusters whose valid and wrong-method traces end in equally long numbers removes this channel, and the separation survives at 0.251 ($[0.151, 0.360]$, $p = 1 \times 10^{-5}$). We treat the shape-matched figure as the headline. It also sharpens the comparison against the unresidualized reduction, which we report throughout: on the full set the residual adds little over raw (0.474 against 0.445), but on the shape-matched set its margin widens (0.251 against 0.196), because the raw signal is the one that answer shape moves. Traces are length-matched (53, 53, 54 median words), so this is not a length effect. The method-versus-execution dissociation reproduces:

the valid-versus-arithmetic-error gap is 0.073 against 0.474 for valid-versus-wrong-method on the same clusters. The arithmetic-error variant is derived programmatically, so its method wording is byte-identical and only the number differs; the residual gap is small but nonzero ($p = 5 \times 10^{-6}$), making R far less sensitive to execution than to method rather than exactly invariant.

The result is not a property of one scorer. Rescoring the same clusters under Mistral-7B, a different family, reproduces the separation with the same sign and the same mechanism: $+0.122$ on the shape-matched subset ($[0.065, 0.186]$, $p = 4 \times 10^{-4}$) and $+0.224$ over all clusters ($p = 1 \times 10^{-15}$), with $\Delta L(\mathcal{P} \mid t)$ falling from 0.497 to 0.241 between a valid and a wrong-method trace while the null-side term barely moves. The magnitude is about half Llama's, which is why we report signs and ranks rather than magnitudes throughout.

A second domain, with no generated data anywhere. Arithmetic still needs a model to phrase the null question; execution tracing does not. Repeating the construction on CRUXEval, where every gold answer comes from running the program and the null is the same program with one operator swapped, the separation holds at $+0.110$ per distinct function ($[0.052, 0.176]$, $p = 6 \times 10^{-6}$, 38 functions) and $+0.174$ under a cross-function null. It is the weakest effect we report and the arm is small, but it is the one where nothing was authored and every leakage and degeneracy channel measures zero (Appendix C.1).

6.2 What residualization buys: invariance, not separation

On the valid-versus-wrong-method contrast the null set does not earn its cost. What residualization buys is invariance to the confound the construct was built around, and the cleanest demonstration is the non-sequitur trace: fluent, on-topic prose that never performs the computation. This is format priming in pure form, and it is exactly the case Theorem 2 says the un-residualized reduction cannot handle. On the same 224 real clusters, the raw reduction *prefers* the non-sequitur to the valid trace (-0.176, that is, it ranks the trace carrying no procedure *above* the one that carries it), while R ranks them correctly ($+0.092$, $p = 9 \times 10^{-7}$; Figure 1b). A signal that inverts on the pure-format case cannot separate reasoning from recitation, however well it performs when the distractor carries no format advantage. On the *templated* benchmark, by contrast, where the non-sequitur is short and carries little format advantage, the raw reduction is the better detector of wrong-method traces (cluster-centered ROC AUC 0.960 against R's 0.889 under Llama-3.1-8B, 0.945 against 0.910 under Mistral-7B). The real-data non-sequitur is fluent and long, which is what a realistic distractor looks like, and there the ordering reverses. So R should not be sold as a stronger detector. Its claim is narrower and matches the theorem: it is the statistic that does not invert when the distractor is well-formed, which is the case that matters for telling reasoning from recitation. It is not confined to one family: under Mistral-7B raw again prefers the non-sequitur (-0.079) and R does not ($+0.030$, not significant). Both signals are reported side by side throughout.

6.3 R is graded with corruption, and separates valid from wrong-method reasoning

Real data cannot vary the corruption of a trace under experimental control, so for that we use the templated benchmark, where a trace is degraded in graded steps with everything else held fixed. Along the semantic ladder (valid, arithmetic-error, wrong-operation, non-sequitur), R falls with severity for every scorer, with severity-R Spearman from -0.385 ($p = 6 \times 10^{-32}$) to -0.071, negative in all five. Structural corruptions preserving the operative computation, sentence shuffling and truncation of the framing, do not lower R; under Llama-3.1-8B they score slightly but significantly *above* valid ($+0.382$, $+0.401$ against $+0.361$; $p = 8.0 \times 10^{-5}$, 1.4×10^{-3}). R tracks the content of the operative step, not its position, which is intended for a measure of method-value but means R is not a trace-quality score: a shuffled trace a human would reject retains the original's value (Figure 2).

Table 2 reports the negative-knowledge contrast $R(\text{valid}) - R(\text{wrong-op})$ on templated traces, where the valid and wrong-operation traces are identical in length and surface form. The separation is robust in relational deduction across all five scorers, with gaps from 0.283 to 0.828 bits per token and p below 10^{-9}. It is strong in arithmetic and list tracing under Llama-3.1-8B, with gaps of 1.267 and 0.276 respectively. Signs replicate across all three data seeds in every cell, and significance in all but one (Qwen2.5-1.5B arithmetic, where one seed reaches $p = 0.06$); magnitudes are stable

Table 1: Faithful-OOD transfer on the independent Mistral-7B model. Mean gain by trace variant, then rank correlations with realized transfer pooled over variants and clusters: R, the raw unresidualized reduction, and R partialling out length, raw reduction, and correctness.

	Mean transfer gain			Predicting transfer			
Domain	valid	wrong-op	acc. (no ex.)	$\rho(R)$	ρ(raw)	partial $\rho(R)$	p
Arithmetic	$+0.054$	-0.762	0.777	$+0.394$	$+0.426$	-0.000	0.99
Deduction	$+0.699$	-0.047	0.297	$+0.799$	$+0.747$	$+0.373$	5.8×10^{-11}
List tracing	$+0.006$	-0.567	0.568	$+0.204$	$+0.149$	$+0.163$	6×10^{-3}
All	$+0.253$	-0.459	0.547	$+0.574$	$+0.575$	$+0.266$	1.6×10^{-15}

in deduction (spread 0.013 to 0.070) but vary where the gap is largest (Llama arithmetic 0.195 on a gap of 1.267), so we emphasize signs and ranks. It is weak or reversed under the Qwen family in arithmetic and list tracing, a pattern the capacity bound below explains. As a check that R is a property of the trace and not a scorer artifact (the second control), we compare scorers on the cluster-centered score. Proposition 2 prescribes this: b is a per-cluster term independent of the trace, so pooling across clusters to compare *traces* carries it as noise. Centered, cross-family agreement is substantial: $+0.419$ for Qwen2.5-7B with Llama-3.1-8B (uncentered $+0.244$), $+0.818$ for Llama-3.1-8B with Mistral-7B ($+0.394$), $+0.623$ for Qwen2.5-1.5B with Llama-3.1-8B ($+0.055$). One of ten pairs falls ($+0.490$ to $+0.384$); we centre because the proposition prescribes it.

Measurement capacity. Corollary 2 predicts a scorer separates traces only insofar as it has posterior mass left to move, for which the baseline codelength on the sibling sets is the observable proxy. Across the fifteen scorer-by-domain cells the mean negative-knowledge separation tracks the mean baseline budget $L_M(\mathcal{P} \mid c_0) + L_M(\mathcal{N} \mid c_0)$ with Spearman $+0.757$ ($p = 1.1 \times 10^{-3}$), and the mean separation stays below the summed budget in every cell (maximum ratio 0.49). The Qwen family is heavily optimized on arithmetic, so its baseline codelength there collapses (to 0.16 to 1.31 bits per token) and it cannot separate; the families that retain budget, Llama-3.1-8B (2.61) and Mistral-7B (0.99 to 1.17), separate with the correct sign (Appendix Figure 4). The relation holds across scorers and domains but *not* across clusters within a single cell, where on the real-data benchmark it reverses (Spearman -0.263, $p = 7 \times 10^{-5}$). The bound therefore tells you which scorer to choose, not which clusters will separate best, and we state it only in that scope.

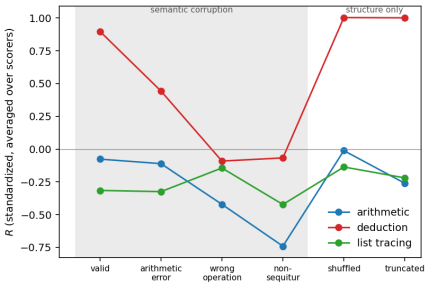

Figure 2: Dose-response of R by domain, standardized within scorer and averaged over the Qwen2.5 and Llama-3.1-8B scorers. Shaded rungs are semantic corruptions of rising severity.

6.4 R PREDICTS REALIZED TRANSFER BEYOND STRONG BASELINES

R predicts realized out-of-distribution transfer beyond trace length, raw reduction, and correctness. The prediction is not circular: R is a codelength under a scorer M, whereas transfer is the generation accuracy of a *different* model M' on a harder held-out set. Realized transfer is governed by method rather than by arithmetic correctness: a trace with the right method but a wrong final number transfers almost as well as a valid trace, whereas wrong-method and non-sequitur traces sharply reduce accuracy. Table 1 gives the mean gains and the rank correlations of scorer-averaged R with transfer. In deduction the partial Spearman controlling for length, raw reduction, and correctness is $+0.373$ ($p = 5.8 \times 10^{-11}$); pooled it is $+0.266$ ($p = 1.6 \times 10^{-15}$). In arithmetic it is near zero, because the raw reduction already captures the available signal there. Figure 5 plots R against realized transfer.

The relation is out of sample: holding out whole clusters in five-fold cross-validation, adding R to length, raw reduction, and correctness raises pooled out-of-sample R^2 from 0.498 to 0.574. Prediction also needs a transfer model with headroom (Appendix C).

A boundary condition: faithful versus compositional shift. The regime of the held-out set matters most to the *un-residualized* signal. Pooled, the marginal correlation with transfer reverses between a faithful and a compositional shift ($+0.574$ to -0.237), but the raw reduction swings harder ($+0.575$ to -0.518), while R's incremental contribution over length, raw reduction, and correctness stays positive in both ($+0.266$ and $+0.226$). Residualization survives the regime change (Appendix C.8).

6.5 R MEASURES METHOD, A PROCESS REWARD MODEL MEASURES STEP-CORRECTNESS

On arithmetic, R and Qwen2.5-Math-PRM-7B are weakly correlated (Spearman $+0.15$) and each predicts transfer beyond the other (partial $+0.286$ and $+0.208$). The cause is a clean dissociation: the PRM scores the arithmetic-error variant lowest because it detects the incorrect step, whereas R scores it as high as a valid trace because the method is preserved. Since transfer is method-governed, the two are complementary rather than redundant (Appendix C.9).

6.6 SCOPE: WHERE R DOES NOT APPLY

The capacity condition above is one boundary. The second concerns what kind of value R tracks: it predicts how much a trace helps a model *in context*, and such traces do not also make better fine-tuning data. From a pool of about twelve chains of thought per cluster, drawn from three teachers and two prompting styles, each selector picks one trace per cluster, so every training set covers the same problems at the same budget and differs only in the trace chosen; we LoRA-tune a Qwen2.5-1.5B student and measure held-out accuracy over three seeds. Selecting the highest-R correct trace is *worse* than a random correct one (0.527 versus 0.583, $p = 0.035$), but this is largely a length artifact, since R anti-correlates with length within the correct pool and terse traces are poor supervision. Selecting within the middle half of that pool by word count halves the deficit and removes its significance (0.562 versus 0.590, $p = 0.30$; versus 0.583 for random-within-correct, $p = 0.47$). The harness detects real effects: correctness filtering beats unfiltered random selection by 0.076. So R carries no usable signal for training-data selection, which rules out the positive effect that would make it a curation tool.

7 DISCUSSION AND CONCLUSION

We studied whether a chain-of-thought trace teaches a transferable procedure or only primes a problem family's format. We proposed R, the codelength reduction a trace induces on problems that need its procedure minus that on format-matched problems that need another, showed that under the mixture view of in-context learning it is the expected log Bayes factor between the two procedures, and found that it separates valid from wrong-method reasoning on real traces, falls with corruption severity, and predicts OOD transfer beyond length, raw codelength, and correctness.

We found that the residual earns its cost on fluent traces that compute nothing. The raw reduction prefers such a trace to the valid one in 76% of real clusters, where R prefers the valid trace in 63% (Figure 1b), although it separates wrong-method traces nearly as well, and our identifiability result shows no transformation of it fixes this. Any method ranking reasoning by how much it lowers a model's loss on related problems thus rewards presentation along with procedure.

We also found that what transfers is the method, not the arithmetic. A trace with the right method and a wrong number transferred almost as well as a valid one and R barely penalized it, yet a process reward model scored it lowest, so correctness filtering discards traces that teach. And a stronger model is not a better instrument. A scorer that already solves the task, as Qwen2.5 does arithmetic, has no codelength left to reduce, and a transfer model at ceiling has no accuracy left to gain.

Contrary to our initial hypothesis, we found the traces R ranks highest are not better fine-tuning data ($p = 0.47$ once length is controlled), so in-context value and training value differ. Our study has further limits. The real-data arms omit deduction, the tracing arm is small, and scoring only the gold answer leaves answer shape as a confound we control by subsampling, which scoring a longer span would remove. Magnitudes are scorer dependent, the process reward comparison is arithmetic-only, and the per-step decomposition does not localize on short traces. Within these limits, R offers a label-free, forward-pass-only way to ask what a reasoning trace actually teaches.

AI USE STATEMENT

Large language models served two distinct roles here, which we separate explicitly. First, as the *object of study*: the scorer, transfer, and generator models (Qwen2.5, Qwen3, Llama-3.1, Mistral-7B, SmolLM3, and the Qwen2.5-Math process reward model) are the systems whose codelength and transfer behaviour we measure, and they produced the real chains of thought, the wrong-method controls, and the rewritten null questions described in Section 5. Every gold answer used for scoring was parsed from a reference solution or obtained by executing code, never generated by a model, and null items were verified against those golds before use. Second, as a *general-purpose assist tool*: large language models were used to help draft and edit the manuscript prose and to help implement and refactor the analysis and plotting code.

ETHICS STATEMENT

This work studies a measurement defined on publicly available open-weight models and on data generated from deterministic symbolic templates; it involves no human subjects, no private or personally identifying data, and no new model training. The intended use of R is to rank or filter reasoning traces by transferable method-value, for example in data curation or distillation. Like any signal read from a scorer model, R inherits the scorer's biases and blind spots, and its validity is bounded by the capacity condition we state and verify; we therefore caution against using it as a standalone quality gate without a non-saturated scorer and a faithful transfer target.

REPRODUCIBILITY STATEMENT

The construction of R is given in Section 2 and Algorithm 1, and all proofs of the theoretical results are in Appendix B. The benchmark, the five scorers, the transfer protocol, and the determinism of the codelength measurement (deterministic templates under three fixed seeds, vLLM prompt log-probabilities in 32-bit precision) are described in Section 5; the concentration, out-of-sample, headroom, per-step, selection, baseline-invariance, and alternate-null analyses are detailed in Appendix C. The templated clusters are generated from deterministic templates under three fixed seeds (42, 123, 456), and the two real-data arms draw their instances from the published GSM-Symbolic and CRUXEval benchmarks under the same three seeds, with every gold answer parsed from a reference solution or obtained by executing the program. Procedure-shifted and null sets are matched by answer-digit count and morphology, and codelength is read from vLLM prompt log-probabilities in true 32-bit precision, so the measurement is deterministic given the scorer. Real chains of thought are sampled at temperature 0.8 and matched wrong-method chains at 0.7; transfer is decoded greedily. Across the significance tests we report, a Benjamini-Hochberg false discovery rate control at $q = 0.05$ leaves every headline result significant. The full experiment and analysis code, the generated principle-clusters, and the scored artifacts are included as supplementary material and are also available at `https://anonymous.4open.science/r/format-residualized-codelength-8D79`, so that reviewers can reproduce the results directly. Every figure and every number in the paper rebuilds from those artifacts on CPU alone, without a GPU and without re-running any scoring: the analysis scripts read the stored codelengths, and the four that would otherwise re-score accept a flag that reads cached per-sibling, per-step, and alternate-null values instead. The supplementary README gives the command for each table and figure, and the generation and scoring stages are provided with their commands for reproducing the pipeline end to end from public datasets and public checkpoints.

REFERENCES

Léonard Blier and Yann Ollivier. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

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

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=9SkkifLopZ.

A. Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–292, 1984. doi: 10.2307/2981683.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *International Conference on Learning Representations*, 2024.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332, 2023.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pp. 5988–6008, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866, 2023.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251, 2019.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Peter D Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=SHMj84U5SH.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pp. 8500–8513, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*, 2021.

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

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 183–196, 2020.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174, 2023.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1819–1862, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025b. doi: 10.18653/v1/2025.findings-acl.547.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50.

Algorithm 1 Format-residualized differential codelength R of a reasoning trace

Require: scorer M; trace t and its seed problem; procedure-shifted set $\mathcal{P}$; format-matched, difficulty-matched null set $\mathcal{N}$; baseline context c_0

1: **for** $\mathcal{S} \in \{\mathcal{P}, \mathcal{N}\}$ and $c \in \{c_0,\, t\}$ **do**
2: $\quad L_M(\mathcal{S} \mid c) \leftarrow$ token-weighted mean over $(x, a) \in \mathcal{S}$ of $-\frac{1}{|a|}\sum_k \log_2 p_M(a_k \mid c, x, a_{<k})$ (one batched forward pass per set; no generation)
3: **end for**
4: $\Delta L(\mathcal{P}) \leftarrow L_M(\mathcal{P} \mid c_0) - L_M(\mathcal{P} \mid t)$ and $\Delta L(\mathcal{N}) \leftarrow L_M(\mathcal{N} \mid c_0) - L_M(\mathcal{N} \mid t)$
5: **return** $R_M(t) \leftarrow \Delta L(\mathcal{P}) - \Delta L(\mathcal{N})$

A Supporting theoretical results

What R measures.

Proposition 1 (Divergence characterization)**.** *The expected per-problem reduction is the reduction in divergence from the true answer law to the model,*

$$\mathbb{E}_z[\delta_t] = \mathbb{E}_x\Big[\mathrm{KL}\big(\mathcal{D}_z(\cdot \mid x) \,\|\, p_M(\cdot \mid x, c_0)\big) - \mathrm{KL}\big(\mathcal{D}_z(\cdot \mid x) \,\|\, p_M(\cdot \mid x, t)\big)\Big], \tag{4}$$

so $R^*_M(t)$ is the difference between the divergence reduction the trace produces on procedure-matched answers and on procedure-mismatched answers.

Equation (4) states that R measures neither plausibility nor correctness but how much closer the trace brings the model to the true answer law on problems that need its procedure than on problems that do not. The contrast is moreover free of the baseline.

How many siblings are needed.

Theorem 3 (Concentration)**.** *If the per-problem reductions $\delta_t(x)$ are independent with $|\delta_t(x)| \le B$, then for $|\mathcal{P}| = |\mathcal{N}| = n$,*

$$\Pr\big(|R_M(t) - R^*_M(t)| \ge \varepsilon\big) \;\le\; 2\exp\big(-n\varepsilon^2/(4B^2)\big),$$

so R_M estimates R^*_M at rate $O(B/\sqrt{n})$, and $n \ge (4B^2/\varepsilon^2)\log(2/\eta)$ siblings give an ε-accurate estimate with probability $1-\eta$.

Appendix C confirms the predicted $1/\sqrt{n}$ rate empirically and explains the near-identical replication of R across the three independent data draws.

R lower-bounds faithful transfer. Call a held-out set $\mathcal{O}$ *faithful* if it is exchangeable with $\mathcal{P}$, that is, drawn from the same procedure z^*.

Theorem 4 (Faithful-transfer identity)**.** *If $\mathcal{O}$ is faithful, then $\mathbb{E}[\Delta L_M(\mathcal{O} \mid t)] = R^*_M(t) + \mathbb{E}[\Delta L_M(\mathcal{N} \mid t)]$. When the trace does not raise the null set's codelength, $\mathbb{E}[\Delta L_M(\mathcal{N} \mid t)] \ge 0$, this gives $R^*_M(t) \le \mathbb{E}[\Delta L_M(\mathcal{O} \mid t)]$, so R lower-bounds the expected faithful-transfer codelength reduction. The side condition is not automatic: it holds for a valid trace in 91.7% of clusters under Llama-3.1-8B and 85.6% under Mistral-7B, but 0.9% under Qwen2.5-7B, whose codelength a trace almost always raises. The bound is thus available exactly on the non-saturated scorers the capacity bound already selects.*

Theorem 4 is the formal content of the residual: a faithful-transfer reduction splits exactly into the procedure-specific part R and a null-set part that is the format priming the residual removes. The bound is in codelength; realized accuracy is a coarsening of it, which is why the empirical relation of Section 6.4 is strong but not deterministic.

Remark 2 (Compositional shift)**.** *If $\mathcal{O}'$ requires a composed procedure $z_O = g(z^*, z')$ with $z' \neq z^*$, exchangeability fails and $\mathbb{E}[\Delta L_M(\mathcal{O}' \mid t)]$ is not tied to $R^*_M(t)$. Under the Bayesian-mixture view of in-context learning (Xie et al., 2022; Saunshi et al., 2021), where $p_M(a \mid x, c) = \int p(a \mid x, z)\, \pi_M(z \mid x, c)\, dz$ and the trace concentrates the posterior on z^*, that concentration removes mass from the concepts supporting z', so a larger R can lower transfer on $\mathcal{O}'$. This is the mechanism behind the boundary condition of Section 6.4.*

B PROOFS

Proof of Proposition 1 (divergence characterization). For any law q, the cross-entropy decomposes as $H(\mathcal{D}, q) = H(\mathcal{D}) + \mathrm{KL}(\mathcal{D} \,\|\, q)$. Substituting into the definition (3) of δ_t cancels the conditional entropy $H(\mathcal{D}(\cdot \mid x))$, which does not depend on the context, and leaves the bracketed KL difference of (4). Taking expectations over the procedure-matched law z^* and the mismatched law z_N and differencing gives $R^*_M(t)$. □

Proof of Theorem 1 (Bayes-factor characterization). Fix a problem x whose answer a is generated by procedure z, and split the mixture at z: $p_M(a \mid x, c) = p(a \mid x, z)\,\pi(z \mid c, x)\,[1 + \rho(a, x, c)]$, where $\rho = \big[\sum_{z' \neq z} p(a \mid x, z')\pi(z' \mid c, x)\big] / \big[p(a \mid x, z)\pi(z \mid c, x)\big]$. Taking negative base-2 logarithms, $\ell(a \mid x, c) = -\log_2 p(a \mid x, z) - \log_2 \pi(z \mid c, x) - \log_2[1 + \rho(a, x, c)]$. Substituting into $\delta_t(x) = \ell(a \mid x, c_0) - \ell(a \mid x, t)$, the term $-\log_2 p(a \mid x, z)$ occurs in both and cancels exactly, with no appeal to separation:

$$\delta_t(x) = \log_2 \frac{\pi(z \mid t, x)}{\pi(z \mid c_0, x)} + \log_2 \frac{1 + \rho_t}{1 + \rho_0}.$$

Under ε-separation both ρ terms are at most ε, so the second term is bounded by $\log_2(1 + \varepsilon)$ in absolute value. Taking expectations over x drawn from z^* and from z_N and differencing gives the claim. □

Proof of Corollary 1 (format cancellation). If $\pi(z \mid t, x) \propto \pi(z \mid c_0, x)$ for every z in the frame, with a proportionality constant depending on the frame and not on z, then the two log-ratios in Theorem 1 are equal and cancel, leaving $O(\varepsilon)$. □

Proof of Corollary 2 (capacity). Rearranging Theorem 1, $R^*_M(t) = \log_2[\pi(z^* \mid t)/\pi(z^* \mid c_0)] + \log_2[\pi(z_N \mid c_0)/\pi(z_N \mid t)]$. The first term is at most $-\log_2 \pi(z^* \mid c_0) = s_0$ because $\pi(z^* \mid t) \leq 1$; the second is at most $\log_2(1/\gamma)$ by the assumption $\pi(z_N \mid t) \geq \gamma\,\pi(z_N \mid c_0)$. □

Proof of Theorem 2 (identifiability). Let t_1 multiply $\pi(z \mid c_0)$ by a common factor $w > 1$ for every z in the frame, and let t_2 multiply only $\pi(z^* \mid c_0)$ by w, leaving the relative masses of the other in-frame procedures unchanged. By Theorem 1 each induces the same expected reduction $\log_2 w$ on $\mathcal{P}$, so $\Delta L_M(\mathcal{P} \mid t_1) = \Delta L_M(\mathcal{P} \mid t_2)$ to $O(\varepsilon)$. By Corollary 1, $R^*_M(t_1) = O(\varepsilon)$, whereas $R^*_M(t_2) = \log_2 w > 0$. Any statistic that is a function of $\Delta L_M(\mathcal{P} \mid \cdot)$ alone therefore assigns t_1 and t_2 the same value and cannot separate them. □

Proof of Proposition 2 (baseline invariance). Substituting (2), $R_M(t) = [L_M(\mathcal{N} \mid t) - L_M(\mathcal{P} \mid t)] + [L_M(\mathcal{P} \mid c_0) - L_M(\mathcal{N} \mid c_0)]$. The second bracket b contains only baseline quantities and is identical for every trace in the cluster, so it cancels in any contrast $R_M(t) - R_M(t')$, which is therefore independent of c_0 and of the $\mathcal{P}$-$\mathcal{N}$ matching that determines b. □

Proof of Theorem 3 (concentration). Write $\mu_P = \mathbb{E}_{z^*}[\delta_t]$ and $\mu_N = \mathbb{E}_{z_N}[\delta_t]$, so $R^*_M(t) = \mu_P - \mu_N$. Then $R_M(t) - R^*_M(t) = \frac{1}{n}\sum_{x \in \mathcal{P}}(\delta_t(x) - \mu_P) - \frac{1}{n}\sum_{x \in \mathcal{N}}(\delta_t(x) - \mu_N)$ is a sum of $2n$ independent, zero-mean terms. Each term has the form $\pm(\delta_t(x) - \mu)/n$ with $|\delta_t(x)| \leq B$, hence range at most $2B/n$. Hoeffding's inequality gives $\Pr(|R_M(t) - R^*_M(t)| \geq \varepsilon) \leq 2\exp(-2\varepsilon^2 / \sum_k (\mathrm{range}_k)^2)$ with $\sum_k (\mathrm{range}_k)^2 = 2n\,(2B/n)^2 = 8B^2/n$, which simplifies to $2\exp(-n\varepsilon^2/4B^2)$. Setting the right side to η and solving for n gives $n \geq (4B^2/\varepsilon^2)\log(2/\eta)$. □

Proof of Theorem 4 (faithful-transfer identity). If $\mathcal{O}$ is exchangeable with $\mathcal{P}$, both are drawn from the procedure z^*, so $\mathbb{E}[\Delta L_M(\mathcal{O} \mid t)] = \mathbb{E}_{z^*}[\delta_t] = \mathbb{E}[\Delta L_M(\mathcal{P} \mid t)]$. By the definition of the population residual, $\mathbb{E}[\Delta L_M(\mathcal{P} \mid t)] = R^*_M(t) + \mathbb{E}[\Delta L_M(\mathcal{N} \mid t)]$. Combining the two equalities yields the identity. When $\mathbb{E}[\Delta L_M(\mathcal{N} \mid t)] \geq 0$, dropping the non-negative second term gives $R^*_M(t) \leq \mathbb{E}[\Delta L_M(\mathcal{O} \mid t)]$. □

Table 2: Negative-knowledge contrast on templated traces, $R(\text{valid}) - R(\text{wrong-op})$ in bits per token, by scorer and domain. Valid and wrong-operation traces are matched in length and surface form; positive values mean R scores valid above fluent wrong-method reasoning. Stars are a one-sided Wilcoxon signed-rank test of $R(\text{valid}) > R(\text{wrong-op})$ over clusters ($^{***}p < 10^{-3}$; every significant cell clears 10^{-3}); the deduction contrast is significant for all five scorers ($p < 10^{-9}$); signs replicate across all three data seeds in every cell and significance in all but one, with magnitude spreads of 0.013 to 0.070 in deduction and up to 0.195 in the largest-gap cells. The reversed Qwen arithmetic and list-tracing cells, hence the non-significant Qwen2.5-7B all-domain cell, are the saturation signature of the capacity bound (Corollary 2); per-cell 95% bootstrap CIs are in Appendix Figure 7.

Scorer	Arithmetic	Deduction	List tracing	All
Qwen2.5-1.5B	$+0.054^{***}$	$+0.751^{***}$	-0.066	$+0.247^{***}$
Qwen2.5-3B	$+0.047^{***}$	$+0.459^{***}$	-0.220	$+0.095^{***}$
Qwen2.5-7B	-0.064	$+0.828^{***}$	-0.249	$+0.172$
Mistral-7B	$+0.135^{***}$	$+0.283^{***}$	$+0.208^{***}$	$+0.209^{***}$
Llama-3.1-8B	$\mathbf{+1.267}^{***}$	$\mathbf{+0.511}^{***}$	$\mathbf{+0.276}^{***}$	$\mathbf{+0.684}^{***}$

C Additional experiments and detail

C.1 The execution-tracing arm

Clusters are CRUXEval functions. $\mathcal{P}$ is the function applied to fresh inputs, with gold answers obtained by executing it; the null is the same function with one operator swapped, executed likewise, and retained inputs are required to make the two programs disagree. Both programs are rendered through the same code formatter, so they differ only in the operator. Traces are recorded from the actual execution, capturing the source line run and the local variables it changed, and both are rendered at a common step budget so they match in length (139 against 143 median words, paired Wilcoxon $p = 0.60$). The wrong-method trace comes from a *second, different* operator edit, so it demonstrates a procedure aligned with neither $\mathcal{P}$ nor the null; an earlier version that reused the null's own edited program inflated the separation by 46 percent, because the wrong-method trace was then a correct demonstration of the null's procedure. Clusters are dropped if a trace states an answer appearing among either set's golds, in either direction, or if a null set's answers are constant. After filtering, all four leakage channels and both degeneracy checks measure zero.

C.2 The two real-data null sets

Figure 3 walks through the construction on one real cluster. The intermediate-quantity null keeps the story fixed and asks for a quantity the reference solution already computes, so the frame matches exactly and the gold answer is parsed from that solution rather than generated; but its required procedure is typically a sub-procedure of the demonstrated one, so the trace genuinely helps it and R is biased downward. The cross-template null takes digit-matched instances of a different template, so no procedure is shared but the surface frame matches only loosely (median Jaccard overlap 0.108), biasing R upward. The overlap the first null carries is measured rather than assumed: under a valid trace it compresses that null 0.300 bits per token more than the cross-template one (95% CI $[0.232, 0.370]$).

C.3 Testing the query-independence idealization

Theorem 1 is stated in expectation over x deliberately. If the posterior were independent of the query, the leading term would be constant across the siblings of a cluster, a strong and testable claim. It is false: in a within-cluster variance decomposition of the per-sibling reduction, the trace explains 0.548 of the variance but the sibling still explains 0.375, and regressing $L_M(x \mid t)$ on $L_M(x \mid c_0)$ across siblings gives a median slope of 0.245 rather than 1 (Appendix C). The exact cancellation of $p(a \mid x, z)$ is untouched by this; only constancy across x is lost, which is why we use the expectation form, and which is all R, a difference of set means, estimates in any case. The slope has its own reading: the reduction grows with an instance's own baseline codelength at a rate below the capacity bound's, the per-instance form of Corollary 2.

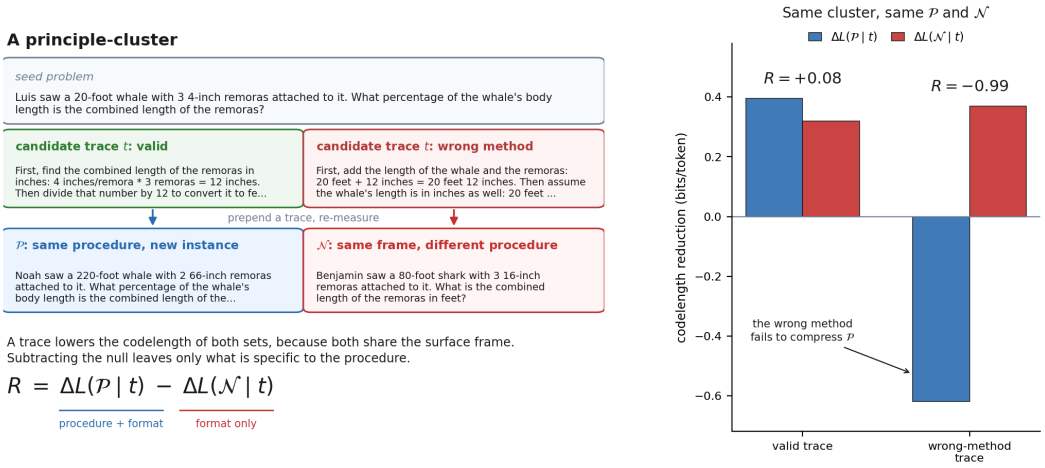

Figure 3: The construction on one real GSM-Symbolic cluster, chosen for legibility; its valid-versus-wrong-method gap is larger than the mean over clusters (Section 6.1). Two candidate traces solve the seed problem, one valid and one using a wrong method, and each is prepended to two sets that share the surface frame but differ in the procedure the question requires: $\mathcal{P}$, other instances of the same procedure, and the null $\mathcal{N}$, whose question needs a different one. The trace lowers the answer codelength of both, because format priming is common to both; the residual R keeps only the procedure-specific part. Right: the same $\mathcal{P}$ and $\mathcal{N}$ under a valid trace and under a wrong-method trace. The wrong-method trace still helps the null, but it *raises* the codelength of the problems that need the demonstrated procedure, so R separates the two sharply.

Proposition 2. *For any cluster,* $R_M(t) = \big[L_M(\mathcal{N} \mid t) - L_M(\mathcal{P} \mid t)\big] + b$*, where* $b = L_M(\mathcal{P} \mid c_0) - L_M(\mathcal{N} \mid c_0)$ *does not depend on* t*. Hence* $R_M(t) - R_M(t')$ *is exactly independent of the baseline context* c_0 *for any two traces* t, t' *in the cluster.*

Proposition 2 is why the negative-knowledge contrast $R(\text{valid}) - R(\text{wrong})$ is a difference-in-differences over the same $\mathcal{P}$ and $\mathcal{N}$: it neither depends on the choice of baseline nor on the quality of the $\mathcal{P}$-$\mathcal{N}$ difficulty matching, since both enter b, which cancels.

C.4 Concentration of R with the number of siblings

Theorem 3 predicts $\text{std}(R) \propto 1/\sqrt{n}$. Resampling siblings within each cluster with replacement and recomputing R for the valid trace, the standard deviation scaled by $\sqrt{n}$ is flat across $n \in \{2, 4, 8, 16, 32\}$ (Table 3), confirming the rate; the constant near 0.31 to 0.34 bits per token implies that the ten siblings we use give $\text{std}(R) \approx 0.10$, consistent with the observed cross-seed spread of the deduction gaps (0.013 to 0.070); the larger spreads in the biggest-gap cells (up to 0.195) exceed this envelope, which is expected since the envelope describes sampling of siblings and not of clusters.

Table 3: Within-cluster subsampling standard deviation of R for the valid trace versus the number of siblings n, averaged over 72 clusters. A flat $\text{std} \cdot \sqrt{n}$ confirms the $1/\sqrt{n}$ rate of Theorem 3.

scorer	n=2	n=4	n=8	n=16	n=32	$\text{std} \cdot \sqrt{n}$
Qwen2.5-7B	0.221	0.157	0.111	0.079	0.055	0.313
Llama-3.1-8B	0.242	0.171	0.122	0.086	0.061	0.343

C.5 Out-of-sample and cluster-level validation

To test that R predicts transfer on unseen clusters, we run five-fold cross-validation in which entire clusters, rather than individual points, are held out, and we predict each held-out trace's transfer gain from a linear model. Adding R to a baseline of trace length, raw codelength reduction, and correctness raises the out-of-sample coefficient of determination from 0.498 to 0.574 pooled over domains,

and improves it in every domain (arithmetic 0.432 to 0.452, relational deduction 0.705 to 0.725, list tracing 0.188 to 0.213). Fitting ordinary least squares of standardized transfer on standardized R, trace length, and correctness, and resampling clusters with replacement, the standardized R coefficient is positive in every domain with 95% intervals excluding zero: $+0.427$ $[+0.317, +0.525]$ in arithmetic, $+0.836$ $[+0.798, +0.881]$ in deduction, $+0.312$ $[+0.197, +0.408]$ in list tracing, and $+0.684$ $[+0.640, +0.723]$ pooled. The arithmetic coefficient is positive once R is not forced to compete with its own collinear raw reduction; the residual's lack of advantage over raw codelength in arithmetic is the partial correlation in Table 1.

C.6 Transfer prediction requires a transfer model with headroom

The capacity bound of Corollary 2 has a counterpart on the transfer side: a method's value can register in realized accuracy only if the transfer model has room to improve and the capability to apply the method. We verify both failure modes with two further transfer models. A model too weak to perform the held-out task, SmolLM3-3B at 0.22 accuracy on the arithmetic OOD set, benefits from any exemplar's format and washes out the method distinction, so R does not predict its transfer (pooled Spearman -0.09). A model already at ceiling, Qwen2.5-7B-Instruct at 0.96 accuracy on the arithmetic and list-tracing OOD sets, leaves no room for an exemplar to help, and again R does not predict (pooled Spearman $+0.08$, not significant). Mistral-7B-v0.3, with intermediate accuracy between 0.30 and 0.78, lies in the band where the method registers and is the transfer model of Section 6.4. The requirement is the same residual-uncertainty condition as for the scorer, now on the evaluation side.

A second in-band model confirms the transfer-prediction result is not specific to Mistral. Llama-3.1-8B, used only as a transfer model and excluded from the scorer-average, has accuracy 0.58 to 0.63 across the three faithful-OOD domains, squarely in the band, and reproduces the method-governed pattern: the valid and arithmetic-error exemplars transfer positively ($+0.42$ and $+0.42$ in deduction, $+0.21$ and $+0.19$ in arithmetic), while the wrong-operation and non-sequitur exemplars transfer negatively. The rank correlation of scorer-averaged R, taken over the four scorers excluding Llama, with Llama's realized transfer is positive in every domain ($+0.247$ arithmetic, $+0.484$ deduction, $+0.103$ list tracing, $+0.318$ pooled, $p = 9 \times 10^{-22}$), and it survives controls for trace length, raw codelength reduction, and correctness pooled ($+0.136$, $p = 6 \times 10^{-5}$) and in arithmetic ($+0.265$, $p = 5 \times 10^{-6}$). Here the raw codelength reduction dominates the residual in deduction, the mirror of the Mistral profile in Table 1, consistent with each model's distinct in-band accuracies; the method-governed transfer signal that R tracks is reproduced on both independent in-band models.

C.7 The per-step overlay does not localize on short traces

Writing $r_i = R_M(t_{\leq i}) - R_M(t_{\leq i-1})$ for the marginal value of step i on the prefix $t_{\leq i}$ through that step, the per-step decomposition does not isolate the operative step on short templated traces. The mean marginal R peaks at the framing or premise step rather than at the operative step, and only 12 to 20 percent of traces have their peak marginal value at the operative step. On these short traces the operation word and the instance-specific arithmetic are entangled within a single step, so the marginal decomposition is noisy; localization plausibly requires longer traces with cleaner step boundaries.

C.8 The faithful-versus-compositional boundary in full

R is calibrated on procedure-shifted siblings, so the regime of the held-out set matters. We report one statistic on both sides rather than switching between them. Pooled, the *marginal* correlation with transfer reverses across regimes, $+0.574$ faithful to -0.237 compositional, but the reversal is carried by the un-residualized signal, whose swing is larger ($+0.575$ to -0.518): under a compositional shift raw compression actively anti-predicts transfer. Controlling for length, raw reduction, and correctness, R's incremental contribution stays positive in *both* regimes, $+0.266$ and $+0.226$; the faithful value is the pooled partial of Table 1, computed on the same scorer average, so the two regimes are compared like for like. So it is the marginal usefulness of codelength that is regime-dependent, while residualization is what survives the change (Figure 6). If the sibling set and the transfer set test different principles, the marginal signal will mislead.

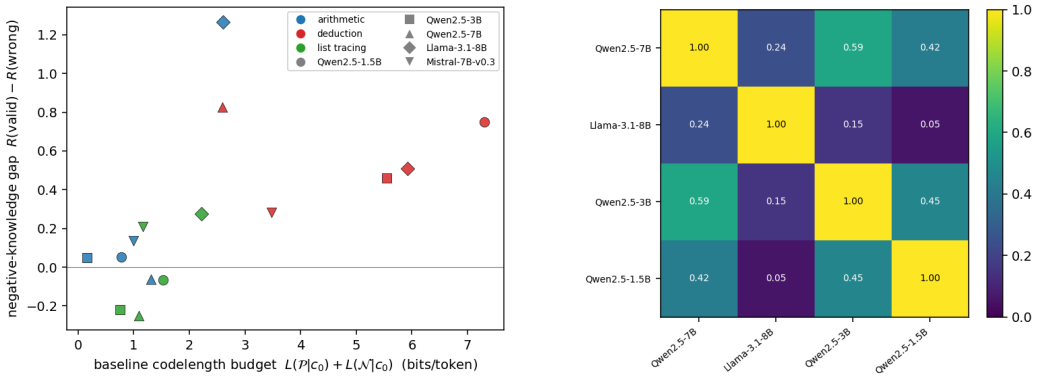

Figure 4: Left: each point is a scorer-by-domain cell, the mean negative-knowledge separation against the mean baseline-codelength budget $L(\mathcal{P} \mid c_0) + L(\mathcal{N} \mid c_0)$; separation tracks the budget (Spearman $+0.757$, $p = 1.1 \times 10^{-3}$), as Corollary 2 predicts. Right: cross-scorer agreement of R (Spearman over clusters and variants), moderate within the Qwen family, lower across families, highest within relational deduction.

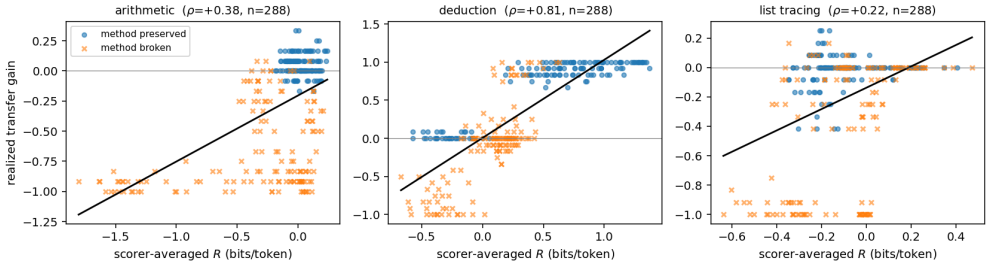

Figure 5: Scorer-averaged R against realized faithful-OOD transfer on the independent Mistral-7B transfer model, over all 216 clusters and four trace variants per cluster, with R averaged over the four scorers excluding the transfer model. Circles are method-preserving traces (valid, arithmetic-error), crosses method-breaking ones (wrong-operation, non-sequitur); the line is an ordinary least-squares fit.

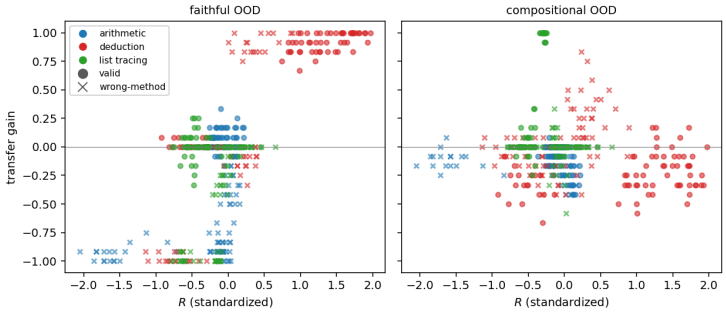

Figure 6: Faithful versus compositional out-of-distribution regimes. The marginal relation between R and realized transfer reverses between the two, while R's incremental contribution over length, raw reduction, and correctness stays positive in both (Section 6.4).

C.9 Process reward model comparison in full

On arithmetic we compare R with Qwen2.5-Math-PRM-7B, with R averaged over Qwen2.5-7B and Llama-3.1-8B. The signals are weakly correlated (Spearman $+0.15$) and each predicts transfer beyond the other (partial $+0.286$ for R given the PRM, $p = 8 \times 10^{-7}$; $+0.208$ the other way, $p = 4 \times 10^{-4}$). Averaging over all four non-transfer scorers instead raises R's partial to $+0.370$,

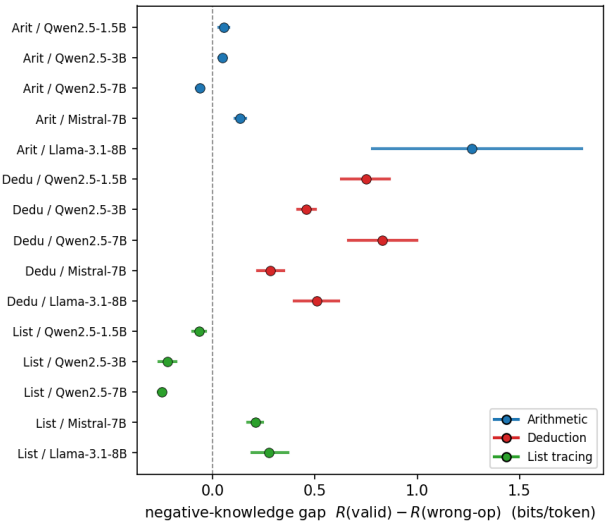

Figure 7: Forest plot of the negative-knowledge contrast $R(\text{valid}) - R(\text{wrong-op})$ for all fifteen scorer-by-domain cells, with 95% bootstrap confidence intervals (10000 cluster resamples). A cell whose interval lies entirely to the right of zero separates valid from wrong-method reasoning: the deduction cells (red) do so for every scorer, and Mistral-7B and Llama-3.1-8B do so in every domain. The cells left of zero are the saturated Qwen arithmetic and list-tracing cells of Table 2, where a scorer that already solves the task reverses rather than separates, as the capacity bound predicts.

so the reported figure is the conservative one. The cause is a clean dissociation: the PRM scores the arithmetic-error variant lowest (0.500 on $[0, 1]$) because it detects the incorrect step, whereas R scores it as high as a valid trace ($+0.30$ versus $+0.35$ standardized) because the method is preserved. Since transfer is method-governed, R aligns with transfer where the PRM does not, so the two are complementary.

C.10 Label-free trace selection

We evaluate R as a label-free selector: among a valid trace and three corruptions, a selector chooses one to supply as the transfer exemplar, scored by that trace's realized faithful-OOD transfer gain on the independent Mistral-7B model (Table 4). Scorer-averaged R matches the transfer oracle in deduction (mean regret 0.000; lift over a random pick $+0.401$, $p = 7 \times 10^{-14}$; a method-preserving trace selected in every cluster). In arithmetic and list tracing the scorer-average is degraded by the saturated Qwen scorers, and the capacity-appropriate non-saturated scorer (Llama-3.1-8B) recovers a positive pooled lift over random ($+0.106$, $p = 0.03$) and over the raw-codelength selector, choosing a method-preserving trace 75% of the time, though it does not reach the label-using oracle or the arithmetic PRM.

Table 4: Label-free trace selection: mean realized faithful-OOD transfer gain of the selected trace (independent Mistral-7B). Among a valid trace and three corruptions a selector picks one exemplar; random is the candidate mean and oracle the best candidate, while pick-valid and PRM use labels or a trained model. R avg is the four non-transfer scorers; R (Llama) is the single non-saturated scorer the capacity bound prescribes.

Domain	random	R avg	R (Llama)	raw-ΔL	PRM	pick-valid	oracle
Arithmetic	-0.302	-0.340	-0.149	-0.207	$+0.054$	$+0.054$	$+0.064$
Deduction	$+0.298$	$\mathbf{+0.699}$	$+0.449$	$+0.699$	–	$+0.699$	$+0.699$
List tracing	-0.211	-0.414	-0.197	-0.512	–	$+0.006$	$+0.015$
All	-0.072	-0.019	$+0.034$	-0.007	–	$+0.253$	$+0.259$

C.11 THE NEGATIVE-KNOWLEDGE CONTRAST IS INDEPENDENT OF THE BASELINE (PROPOSITION 2)

The absolute score $R(\text{valid})$ correlates with the baseline term $b = L_M(\mathcal{P} \mid c_0) - L_M(\mathcal{N} \mid c_0)$ (Pearson $+0.43$ to $+0.82$ across the five scorers), since $R(t) = b + [L_M(\mathcal{N} \mid t) - L_M(\mathcal{P} \mid t)]$. The negative-knowledge contrast $R(\text{valid}) - R(\text{wrong-op})$ is, by contrast, bit-identical to within 10^{-15} bits whether computed with the baseline, from the trace-conditioned codelengths alone, or after adding an arbitrary per-cluster shift to c_0, an exact empirical confirmation that contrasts of R are c_0-invariant (Proposition 2).

C.12 ROBUSTNESS TO THE CHOICE OF NULL SET

The contrast also does not depend on the particular null set. We build an alternate null $\mathcal{N}'$ for every cluster by rotating each principle to a different mismatched principle, distinct from the original null and digit- and morphology-matched to the same $\mathcal{P}$, re-score $\mathcal{N}'$ under Llama-3.1-8B in 32-bit precision, and recompute the negative-knowledge contrast (reusing the original $\mathcal{P}$-side codelengths). The alternate-null contrast tracks the original almost exactly in arithmetic ($+1.274$ versus $+1.267$ bits per token, per-cluster Pearson 1.00), strongly in deduction ($+0.343$ versus $+0.511$, Pearson 0.93, Wilcoxon $p = 6 \times 10^{-5}$) and pooled over domains ($+0.527$ versus $+0.684$, Pearson 0.98, $p = 1.4 \times 10^{-8}$); only in list tracing, the domain with the smallest original contrast, does the alternate null fail to preserve significance. The separation of valid from wrong-method reasoning is therefore a property of the trace and its procedure-shifted siblings, not of the specific null.