When Quality Filters Backfire: Agreement-Based Latent Filtering Hurts Bootstrap Pretraining

Suan Lee **JaeSeong Kim**
School of Computer Science, Semyung University
{suanlee, kjsqp1010}@semyung.ac.kr

Abstract

Latent thought pretraining augments training text with model-generated "thoughts". Because generating them is expensive, the obvious way to scale is to spend the budget only on "high quality" chunks—and the most intuitive quality signal is whether the thought generated *before* a chunk agrees with the one generated *after*. We show this intuition backfires. In a controlled 1.1B experiment that compares every allocation rule with a random baseline at its *own* inclusion rate, keeping only high-agreement chunks raises held-out perplexity above the budget-matched random baseline in every seed ($+0.84\%$; n=5 vs. 3, exact permutation p=0.018). The deficit persists across a second training iteration, though downstream accuracy does not separate the filter from random at the same budget. We trace the cause: importance rises with agreement over most of its range, but the top agreement decile, where prior and posterior coincide or the prior is missing, holds degenerate latents that an agreement-only filter cannot exclude. To diagnose this we introduce PAIR, which pairs agreement with a length-corrected importance weight (raw importance is itself a length artifact, $\rho=-0.786$). Importance routing alone matches random at its budget, the intersection of the two filters is slightly *worse* than random ($+0.3\%$, p=0.008), and inverting the importance signal collapses perplexity. Budget matching decides these verdicts: against a random baseline at a lower inclusion rate the agreement filter looks 4.9% worse and importance routing 1.2% worse, and an earlier version of this work credited the intersection with a gain that was likewise a budget effect. Intuitive single-axis quality filters do not beat random allocation, and any latent allocation study must compare each rule with a random baseline at its own inclusion rate.

1 Introduction

Latent thought pretraining augments raw text with model-generated latent "thoughts" before training, and consistently yields perplexity improvements over standard next-token prediction (Ruan et al., 2025; Wang et al., 2025; Zhu et al., 2025; Yang et al., 2025; Hao et al., 2024; Kong et al., 2025). Generating those latents is expensive, so a natural way to scale the paradigm is to spend the latent budget only on the chunks where it pays off.

An intuitive but harmful quality filter. Data-selection methods routinely keep only the "high quality" fraction of a corpus (Xie et al., 2023a,b; Albalak et al., 2023), and the natural quality signal for a self-generated latent is the model's own *self-consistency*—the same principle that underlies bootstrapping reasoning from a model's agreeing samples (Zelikman et al., 2022, 2024). Read off a BOLT-style pipeline, that signal is *prior–posterior agreement*: each chunk carries a planning latent z^{pri} drawn before the chunk and a reflecting latent z^{post} drawn after; if the two agree, the model appears to have a coherent, confident representation of the chunk, so keeping only the high-agreement chunks should retain "the good latents". We test this filter in a controlled experiment (1.1B-parameter math pretraining, n=5 seeds per allocation rule, each compared with a random baseline re-rolled from the same latents at the rule's own inclusion rate) and find the opposite: the agreement-only condition (*Bridge*) is worse than its budget-matched random baseline in every seed (val PPL $+0.84\%$, exact permutation p=0.018; Table 1). The deficit persists into a second EM iteration (Section 5.5); downstream, Bridge's low GSM8K and MATH scores are matched by random allocation at the same budget (Section 5.4).

Why the filter fails. We interpret this as a Goodhart-style failure. Let IG_i be the information gain that observing the suffix delivers to the posterior (formally a KL between posterior and prior, §5.3). The intuitive worry is that high agreement means the suffix barely moved the posterior, so the

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2672–2687
November 6-10, 2026 ©2026 Association for Computational Linguistics

latent carries little new information. Measured on 2.46M chunks the picture is sharper (§5.3): importance *rises* with agreement through the ninth decile and then falls in the top decile, which mixes genuine posterior collapse with chunks whose prior is missing and is recorded as perfect agreement. An agreement-only filter keeps exactly that tail.

A diagnostic and its predicted single-axis harm. Diagnosing this requires a second axis. We introduce PAIR (**P**rior-posterior **A**greement + **I**mportance **R**outing), pairing $\mathcal{A}(x)$ with a length-corrected per-token importance weight $\tilde{\mathcal{I}}(x) = \mathcal{I}(x)/|x|$. The raw $\mathcal{I}(x) = \log p_\theta(x \mid z) - \log p_\theta(x)$ used by prior analyses is itself a length artifact ($\rho(\mathcal{I}, |x|) = -0.786$, $p < 10^{-10}$, n=11.8k); after controlling for length and entropy, $\tilde{\mathcal{I}}$ and $\mathcal{A}$ are conditionally independent (partial $\rho = -0.011$, $p = 0.25$). The three rules keep different fractions of chunks (82%, 50%, 41%), so each is compared with a random baseline at its own inclusion rate; fewer latents alone lower random-routing PPL, so an unmatched baseline can manufacture or hide an effect (§5.2). At matched budget, Bridge is worse than random, importance routing (Route) is indistinguishable from it, and their intersection (Combined) is slightly *worse* than random—so the earlier draft's "combine the two filters" remedy was only a budget effect, and we retract it.

Contributions.

- **Two negative results with a mechanism.** At matched budget, agreement-only filtering is worse than random in every seed (val PPL +0.84%, exact permutation p=0.018), and reversed-importance routing degrades val PPL by +15.0% over a coverage-matched random control. The Bridge val-PPL deficit persists across a second EM iteration; downstream, Bridge is within seed noise of random at its budget. We give a Goodhart/information-gain hypothesis (§5.3) with three enumerated predictions (P1, P2, P3), each corroborated by an independent experiment.
- **A length artifact in prior importance analyses.** Raw importance weight correlates with chunk length at $\rho = -0.786$ ($p < 10^{-10}$, n=11.8k chunks), so "high-importance" rankings partly select short chunks; the corrected $\tilde{\mathcal{I}}$ is conditionally independent of agreement (partial $\rho = -0.011$).
- **PAIR, a two-axis diagnostic, tested against budget-matched random baselines.** Each rule is compared with random allocation at its own inclusion rate. Route is indistinguishable from random (-0.09%, p=0.73), and the intersection (Combined) is significantly *worse* than random ($+0.33\%$, p=0.008); the advantage Combined held over an unmatched Random@50% baseline was entirely a budget effect, so we *retract* the earlier intersection-as-remedy claim. The transferable lesson is that latent-allocation studies must compare each rule with a random baseline at its own inclusion rate.

Terminology. *Agreement*: overlap between a chunk's before-text (prior) and after-text (posterior) latents; *importance*: how much the latent raises the model's likelihood of the chunk (length-corrected, §3.2). The allocation *conditions* decide which chunks keep a latent—*Bridge* (high-agreement only), *Route* (high-importance only), *Combined* (their intersection), *Random* (a random subset), and *Reversed* (an ablation keeping the *lowest*-importance chunks). A *matched-budget* comparison fixes the fraction of chunks that keep a latent, so differences reflect the *criterion*, not the amount of latent text. "Goodhart-via-extremes" names the failure in which pushing a proxy to its extreme selects degenerate examples. Full notation: Appendix A.

2 Background

Latent thought pretraining. Latent thought pretraining casts language modeling as inference over a latent variable model: given text x, a latent z represents implicit reasoning, and the model maximizes the ELBO $\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(z|x)}[\log p_\theta(x \mid z)] - \text{KL}[q(z \mid x) \,\|\, p_\theta(z)]$. BOLT (Ruan et al., 2025) implements this as an EM loop: the E-step samples latents from $q_\theta(z \mid x)$ and the M-step retrains on (z, x) pairs, a self-improving cycle. TPT (Wang et al., 2025), OURO (Zhu et al., 2025), and SBP (Yang et al., 2025) vary the parameterization; COCONUT (Hao et al., 2024) and LTM (Kong et al., 2025) address the fine-tuning regime. PAIR is defined over BOLT but applies to any method producing both prior and posterior latent streams.

Prior and posterior channels. BOLT associates each chunk x with a **prior latent** z^{pri} generated *before* observing x (a "planning" signal) and a **posterior latent** z^{post} generated *after* (a "reflecting" signal). The raw **importance weight** and its per-

token **normalized** form are

$$\mathcal{I}(x, z) = \log p_\theta(x \mid z) - \log p_\theta(x), \quad \tilde{\mathcal{I}} = \frac{\mathcal{I}}{|x|}, \tag{1}$$

and the **prior–posterior agreement** is $\mathcal{A}(x) = \text{sim}(f(z^{\text{pri}}), f(z^{\text{post}}))$, the token-level F1 overlap between the two latent sequences. High $\mathcal{A}$ means planning and reflecting signals converge.

3 PAIR: Prior-posterior Agreement + Importance Routing

3.1 The Length Artifact Problem

A fundamental issue with using raw importance $\mathcal{I}(x, z)$ as a routing signal is that it is an *unnormalized sum* over token positions. For a chunk with $|x|$ tokens, the expected raw importance is:

$$\mathbb{E}[\mathcal{I}(x, z)] \approx |x| \cdot \bar{\imath}(x) \tag{2}$$

where $\bar{\imath}(x)$ is the mean per-token importance. Even if $\bar{\imath}(x)$ is constant across chunks, $\mathcal{I}$ scales linearly with length. This induces a mechanical negative correlation with length *when ranking by importance*, because high-importance-ranked chunks tend to be short (their per-token signal dominates the ranking).

Figure 1(a) shows the scatter of raw importance vs. length on FineMath-4+ data (an n=11,859 importance-quartile stratified subsample of the 2.46M cache used in §5.3). We observe $\rho = -0.786$ ($p < 0.001$), confirming that raw importance is a poor proxy for latent benefit per unit compute.

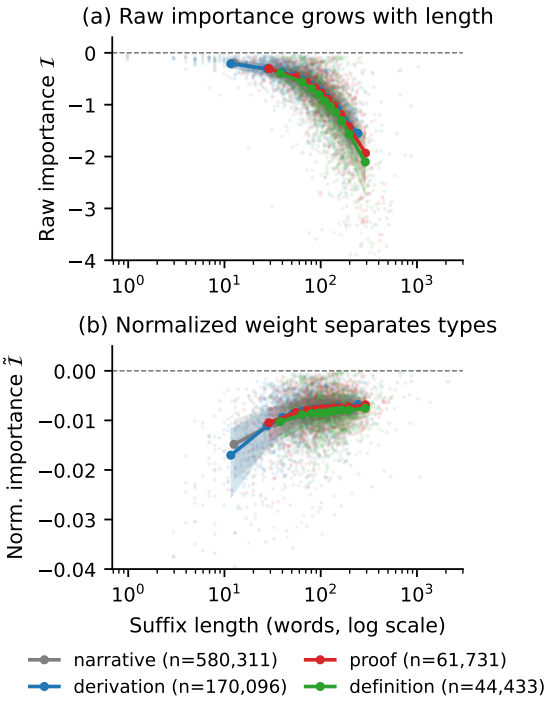


Figure 1: Raw importance weight is strongly confounded with chunk length (panel a; Spearman $\rho = -0.786$, $p < 10^{-10}$ on 11,859 chunks), while the normalized importance weight $\tilde{\mathcal{I}} = \mathcal{I}/|x|$ removes this artifact by construction (panel b). Lines are per-type medians over length deciles (bands: interquartile range); points are coloured by text type. After normalization, $\tilde{\mathcal{I}}$ means separate modestly across text types (see Section 3.2).

3.2 Normalized Importance Weight

Dividing by token count (Equation 1) removes the linear length scaling that inflates raw $\mathcal{I}$: over the full 2.46M-chunk cache the rank correlation with suffix length moves from ρ=−0.72 (raw) to ρ=+0.29 ($\tilde{\mathcal{I}}$), a weak residual trend visible in Figure 1(b). Aggregated over 856,571 math-domain chunks produced by the *combined* generation condition, per-type means of $\tilde{\mathcal{I}}$ are indeed modestly separated (proof −0.0083, definition −0.0087, narrative −0.0104, derivation −0.0105; higher is better because importance weights are negative log ratios), with standard deviations $\approx$ 4–13 $\times 10^{-3}$. The effect sizes are small (Cohen's $d \approx 0.18$ for proof vs. narrative) but consistent across random seeds of the chunking process. We therefore report a *directional* per-type ordering rather than a strong claim that any single type dominates.

An allocation tension. Per-type inclusion rates under the *combined* condition expose a subtler issue. Although proofs and definitions have the *least* negative $\tilde{\mathcal{I}}$ (i.e. the best per-token latent benefit), they receive the *lowest* routing inclusion (33.7% and 26.0% respectively). Derivations show the *most* negative $\tilde{\mathcal{I}}$ yet are routed most aggressively (48.8%). The inversion is driven by residual length dependence of the routing score prior to our per-suffix normalization fix (§3.2): derivation suffixes are short (mean 91 words) relative to proof (127) and definition (134) suffixes, and the generation-time routing threshold was computed using a latent-length proxy rather than suffix length. The partial-correlation fix neutralises this effect in *analysis*, but the factorial experiments below reuse the original generation artifacts, so we report this mismatch as a known limitation of the routed samples used here.

3.3 Prior-Posterior Agreement as Independent Signal

Agreement $\mathcal{A}(x)$ captures a qualitatively different property: whether the model's planning and reflecting latents *converge*. The raw Spearman ρ between $\mathcal{I}$ and $\mathcal{A}$ is -0.060 ($p < 10^{-10}$), reflecting a shared length confound; after controlling for chunk length, lexical entropy, and character entropy, the partial ρ collapses to -0.011 ($p = 0.25$, n.s.), confirming the two axes are conditionally independent. The two axes induce four quadrants, which we name with neutral labels to avoid confusion with the experimental *conditions* (Bridge/Route/Combined/Reversed) used later:

	High $\mathcal{A}$	**Low** $\mathcal{A}$
High $\tilde{\mathcal{I}}$	Core	Diverse
Low $\tilde{\mathcal{I}}$	Stable	Omit

Core chunks (high on both) are informative and coherent. *Diverse* chunks are informative but exploratory. *Stable* chunks are coherent but low-gain. *Omit* chunks are candidates for no augmentation. The experimental conditions map onto this space: the *Bridge* condition filters on $\mathcal{A}$ alone (Core $\cup$ Stable), *Route* on $\tilde{\mathcal{I}}$ alone (Core $\cup$ Diverse), and *Combined* on their intersection (Core). The hypothesis of §5.3 is that Core is the productive region and that filtering on $\mathcal{A}$ alone fails because it admits the low-gain Stable quadrant.

3.4 Adaptive Allocation via Bridge + Route

Given PAIR scores from bootstrap iteration N, we allocate latent budget in iteration $N{+}1$ as follows. Chunks above the k-th percentile of $\tilde{\mathcal{I}}$ are selected as *route* candidates; among those, chunks above the m-th percentile of $\mathcal{A}$ are promoted to *bridge*. Remaining chunks receive `null` (no latent augmentation).

This design targets a reduced latent token budget; the resulting end-to-end compute trade-off relative to uniform allocation is quantified in Section 5.6.

3.5 Statistical Methodology

All Spearman correlations use 95% bootstrap confidence intervals (1000 resamples). Each allocation rule (seeds 42–46) is compared with a random control re-rolled from the same latents at the rule's own inclusion rate: Random@82 for Bridge (n=3), Random@50 for Route (n=6), and Random@41 for Combined (n=5). For each contrast we report Welch's t-test and the exact two-sided permutation p over all seed relabellings, whose floor depends on the seed counts (0.008 for 5 vs. 5); the full pairwise grid is in Appendix B. Our primary endpoint is model-level held-out perplexity, as distinct from chunk-level per-example metrics. The Reversed ablation and its coverage-matched control have n=3 each (directional only).

4 Experimental Setup

4.1 Data and Model

Datasets. Our primary experiments use **Math**: FineMath-4+ (documented in Allal et al., 2025), high-quality mathematical web text filtered to quality score ≥ 4. The full bootstrap split contains approximately 1.92B raw tokens; we use a 20% subsample (approximately 384M raw tokens and 1.3B total tokens after latent augmentation) for the factorial comparisons reported here. A second domain, FineWeb-Edu (Penedo et al., 2024) educational web text, is used for the replication in §5.7. Train/validation splits follow the BoLT reference split (`finemath_4plus_final_fixed_val_split`).

Model. We use TinyLlama-1.1B (Zhang et al., 2024) as the base transformer (22 layers, hidden dimension 2048, 32 attention heads, $\sim$1.1B parameters). Latents are generated (E-step) by a warmstart checkpoint trained for 4,069 optimization steps ($\sim$800M tokens, including GPT-generated teacher latents); every M-step run starts from the pretrained TinyLlama-1.1B checkpoint.

Training. Each PAIR condition is trained for 3,306 optimization steps ($\sim$1 epoch over the 1.3B-token augmented corpus: $192 \times 2048 \times 3{,}306 \approx 1.30$B) with global batch size 192 (per-GPU batch 12, gradient accumulation 4 across 4 A100 GPUs), sequence length 2048, cosine learning-rate schedule (peak 1×10^{-4}, 1,000 warm-up steps), and five random seeds per confirmatory condition (§3.5).

Latent generation. Latents are generated with vLLM (Kwon et al., 2023) at temperature $\tau = 1.0$ and maximum length 1,000 tokens. For each chunk we draw $K = 4$ posterior samples and select one via Monte-Carlo importance weighting; for Bridge/Combined conditions we additionally generate a prior sample to compute $\mathcal{A}$. Chunking splits documents by sentence with a Poisson-distributed unit count (mean 8 words, max 20).

4.2 Conditions

We evaluate three allocation rules against uniform random allocation, plus a reversed-routing ablation

(selection rules and measured inclusion rates in Table 1): *Bridge* keeps chunks with $\mathcal{A} \geq 0.3$ (82.3% of chunks), *Route* keeps the top 50% by $\tilde{\mathcal{I}}$ (50.0%), and *Combined* intersects the two filters (41.2%). The filters do not produce equal budgets, so each is compared with a random baseline re-rolled from the same latents at its *own* inclusion rate: Random@50 (50.1%) for Route, Random@41 (41.4%) for Combined, and Random@82 (82.3%) for Bridge (§5.2). Bridge has high inclusion because the agreement threshold is calibrated for filtering out only clearly low-agreement chunks. Two further conditions (uniform 100% and zero-latent baseline) are deferred to follow-up experiments.

4.3 Evaluation Protocol

Our primary endpoint is held-out perplexity on the FineMath-4+ validation split (1,000 samples). We additionally report three downstream benchmarks (GSM8K, MATH, MMLU; §5.4) and, to check for latent collapse across EM iterations, latent-diversity metrics (vocabulary size, unique-4-gram rate, self-BLEU-4) on a fixed 1,500-latent sample per (condition, iteration), reported in §5.5 and Appendix B.

5 Results

5.1 The Importance Weight Artifact

Figure 1 confirms the length artifact derived in §3.1: raw importance is strongly length-confounded ($\rho = -0.786$), and the per-token normalization removes it. We use $\tilde{\mathcal{I}}$ for all allocation decisions below; per-text-type $\tilde{\mathcal{I}}$ statistics (definition/proof/derivation/narrative) are deferred to Appendix E as they are not load-bearing for the factorial results.

5.2 The Agreement Filter Backfires: Factorial Validation

We trained each allocation rule for 3,306 optimization steps (∼1 epoch over the 1.3B-token augmented corpus) with five seeds (42–46), together with a random control at each rule's inclusion rate: Random@50 (n=6), Random@41 (n=5) and Random@82 (n=3), plus the Reversed ablation and its coverage-matched control (n=3 each). After training we measured held-out validation perplexity on 1,000 FineMath samples. Table 1 summarises the results, and Figure 2 plots them. The headline result is the Bridge row: filtering by prior–posterior agreement alone is worse than random allocation at the same budget.

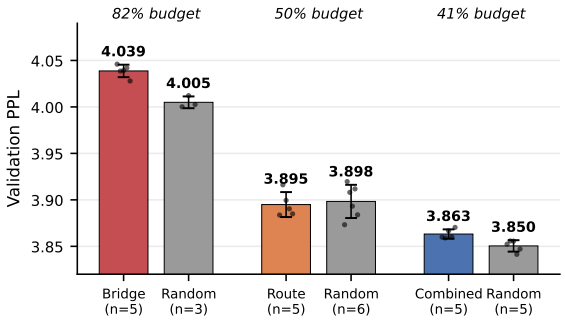


Figure 2: Validation PPL (mean ± std; dots are seeds), each allocation rule next to the random control at its own inclusion rate. Bridge is worse than Random@82 in every seed (+0.84%, permutation p=0.018), Route matches Random@50 (−0.09%, p=0.73), and Combined is worse than Random@41 (+0.33%, p=0.008). Across the random controls fewer latents give lower PPL, so comparing a rule with random at a different budget mixes in this budget effect. See Table 1.

Four findings. (i) Bridge is worse than random at its own budget. Bridge (82.3% inclusion) is worse than Random@82 (82.3%) in every seed: val PPL +0.84% (Welch p=0.001; exact permutation p=0.018, the smallest value attainable with 5 vs. 3 seeds). The gap is far smaller than against the lower-budget baselines (+3.60% vs. Random@50, +4.89% vs. Random@41), most of which is the budget effect itself: Random@82 sits +2.74% above Random@50. The deficit persists into iter 2 (§5.5). **(ii) Route is indistinguishable from Random at its own budget.** Route keeps 50.0% of chunks, so its budget-matched baseline is Random@50 (50.1%), against which it is neutral ($\Delta = -0.09\%$, $p = 0.73$). Its +1.16% gap to Random@41 is a budget difference, not a routing effect: Random@50 itself sits +1.24% above Random@41. Importance-percentile routing alone thus neither helps nor hurts val PPL. **(iii) Combined is *not* a remedy: it is significantly worse than the matched-budget Random.** Combined's −0.90% apparent advantage over Random@50 is fully explained by the budget difference: Random@50→41 alone gains −1.23%. Against the budget-matched Random@41, Combined is significantly *worse* by +0.33% (Welch p=0.008; exact permutation p=0.008). We therefore retract any claim that the $\mathcal{A}$-$\tilde{\mathcal{I}}$ intersection provides routing benefit; at matched budget the two-axis filter is a (small) net harm, not a gain.

(iv) Reversed routing is a directional sign-test on $\tilde{\mathcal{I}}$. Allocating latents to the *lowest*-$\tilde{\mathcal{I}}$ chunks de-

Table 1: Validation perplexity after iteration 1 (held-out FineMath-4+ split, 1,000 samples; lower is better), grouped by inclusion budget. *Latent %* is the measured fraction of training chunks carrying a latent. Each rule is compared with a random control re-rolled from the same latents at its *own* inclusion rate: $\Delta\%$ is the relative gap in mean PPL, Welch p is two-sided, and Perm. p is the exact two-sided permutation p (floor 0.008 for 5 vs. 5, 0.002 for 5 vs. 6, 0.018 for 5 vs. 3, 0.10 for 3 vs. 3). Bold: significant under both tests. †Trained on a 6-slice corpus subset, with a random control on the same slices; Reversed carries fewer latents than that control, which favours Reversed, so its $+15.0\%$ gap is conservative.

				Val PPL		vs. budget-matched Random				
Condition	**Selection rule**	n	**Latent %**	**Mean ± std**	**Seed range**	**Control**	**Δ%**	**Welch p**	**Perm. p**	**Δ% vs R@50**
≈82% budget										
Bridge	$\mathcal{A} \geq 0.3$	5	82.3%	4.039 ± 0.007	4.028–4.046	Random@82	**+0.84**	0.001	0.018	+3.60
Random@82	uniform	3	82.3%	4.005 ± 0.006	4.000–4.012	–	–	–	–	+2.74
50% budget										
Route	top-50% $\tilde{\mathcal{I}}$	5	50.0%	3.895 ± 0.013	3.884–3.916	Random@50	−0.09	0.729	0.727	−0.09
Random@50	uniform	6	50.1%	3.898 ± 0.018	3.873–3.920	–	–	–	–	0.00
≈41% budget										
Combined	Bridge ∩ Route	5	41.2%	3.863 ± 0.005	3.859–3.870	Random@41	**+0.33**	0.008	0.008	−0.90
Random@41	uniform	5	41.4%	3.850 ± 0.006	3.841–3.856	–	–	–	–	−1.23
6-slice corpus (coverage-matched ablation)										
Reversed†	bottom-50% $\tilde{\mathcal{I}}$	3	40.1%	5.100 ± 0.014	5.087–5.114	Random†	+15.01	$<10^{-4}$	0.100	–
Random†	uniform	3	50.2%	4.435 ± 0.019	4.418–4.455	–	–	–	–	–

grades val PPL in the predicted direction (Table 1; raw $\Delta = +30.8\%$ vs. Random@50). A coverage-matched control (random routing on the identical 6-slice subset) isolates the effect: Reversed stays $+15.0\%$ above matched-coverage Random, the rest being corpus size (Limitations), even though it carries fewer latents (40.1% vs. 50.2%). Inverting $\tilde{\mathcal{I}}$ is by far the most harmful allocation, evidence that $\tilde{\mathcal{I}}$ is correctly signed (§5.3).

The train-loss ranking (Combined $<$ Route $\ll$ Bridge $\approx$ Random@50; App. B) does not track val PPL. Training loss mixes latent and text tokens in proportions that differ by inclusion budget, so we do not read it as evidence for or against any rule.

5.3 Why Bridge Filtering Hurts: A Measured Goodhart Failure

Define the per-chunk *information gain* as the KL between the posterior (suffix-aware) and prior (suffix-blind) latent distributions,

$$\mathrm{IG}_i = \mathrm{KL}\big(q_\theta(\cdot \mid x_i^{\text{pre}}, x_i^{\text{suf}}) \,\big\|\, p_\theta(\cdot \mid x_i^{\text{pre}})\big), \quad (3)$$

the amount the suffix moves the posterior. Via the ELBO $\log p_\theta(x) \geq \mathbb{E}_q[\log p_\theta(x \mid z)] - \mathrm{IG}$, the per-token importance weight $\tilde{\mathcal{I}}$ is a tractable surrogate for IG. We measure the empirical $\mathcal{A} \leftrightarrow \tilde{\mathcal{I}}$ relationship directly at the *chunk* level on the full cached corpus (n=2,455,877 chunks; Figure 3).

Non-monotone, with a degenerate high-agreement tail. A naive hypothesis—$\mathcal{A}$ *monotonically* decreasing in IG—is falsified: $\mathcal{A}$ and $\tilde{\mathcal{I}}$ are weakly *positively* rank-correlated overall (ρ=+0.17). Instead the relationship is *inverted-U*: $\tilde{\mathcal{I}}$ rises monotonically with agreement through the 9th decile ($-13.7 \to -8.3 \times 10^{-3}$) then *falls* in the top decile back to the sixth-decile level (-9.6×10^{-3}). The extreme high-agreement tail is worth markedly less than the ninth decile. Two populations sit there: chunks whose prior and posterior genuinely near-coincide (posterior collapse), and—materially—chunks for which the E-step emitted no prior at all, which the cache records with a default agreement of 1.0. Both are degenerate for one practical reason: an agreement-only filter cannot distinguish a confidently redundant latent from an unmeasured one. This is a *Goodhart-via-extremes* failure: agreement is sensible in the bulk but its *maximum* marks a corner that mixes trivial agreement with missing measurement. **This is not a length artifact**: D9 and D10 have nearly identical mean suffix length (93 vs. 99 words) yet $\tilde{\mathcal{I}}$ drops $-8.3 \to -9.6 \times 10^{-3}$, and the rank-correlation stays positive within short/mid/long length tertiles ($+0.16/+0.15/+0.19$). The same inverted-U holds within all four math text types (App. B, Fig. 4). The pattern is also not a token-F1 artifact: an embedding-cosine agreement (`all-MiniLM-L6-v2`; Reimers and Gurevych, 2019; 50k chunks) correlates with token-F1 at Spearman ρ=0.61 and reproduces the decile pattern.

Three predictions, each checked by an independent experiment. The measured inverted-U yields three falsifiable consequences:

- **(P1) Bridge harm.** An $\mathcal{A}$-only filter (≥ 0.3, 82.3%) cannot exclude the degenerate top decile, so it should underperform Random at the same budget. *Checked* (§5.2): $\Delta{=}{+}0.84\%$ vs. Random@82, worse in every seed (exact permutation $p{=}0.018$), but the effect is small.
- **(P2) $\tilde{\mathcal{I}}$ is correctly signed.** Routing latents to the *lowest-*$\tilde{\mathcal{I}}$ chunks should be worse than $\mathcal{A}$-only filtering. *Checked* (§5.2): Reversed produces the largest swing in the study, in the predicted direction; a coverage-matched control isolates a $+15.0\%$ reversed-routing effect (Limitations).
- **(P3) The deficit is persistent.** A selection-induced Goodhart failure (distinct from diversity collapse) should remain stable across bootstrap iterations. *Checked* (§5.5): Bridge moves only $+0.005$ PPL across iters $1 \to 2$, no diversity collapse.

The intersection (Combined) would, on a naive reading, escape both the degenerate top-$\mathcal{A}$ tail and the low-$\mathcal{A}$ bottom. The matched-budget control (Random@41, §5.2) refutes this: Combined's -0.90% gap vs Random@50 is fully explained by the inclusion-budget difference, and vs the budget-matched Random@41 Combined is significantly *worse* ($+0.33\%$, $p{=}0.008$). We therefore treat (P1, P2, P3) as the empirically supported diagnostic content and do *not* claim an intersection benefit. The full decile curve is in App. B (Fig. 3).

5.4 Downstream Tasks: No Budget-Matched Deficit

We evaluate the 18 iter-1 checkpoints (six conditions, seeds 42–44) on three benchmarks via `lm-eval-harness` (Gao et al., 2024): GSM8K (Cobbe et al., 2021) (8-shot chain-of-thought, CoT (Wei et al., 2022), with both *strict-match* and *flexible-extract* graders), Hendrycks MATH (Hendrycks et al., 2021b) (4-shot CoT), and MMLU (Hendrycks et al., 2021a) (synthetic CoT format).[1] Table 2 reports the mean ± std ($n{=}3$).

[1]MMLU is an *out-of-domain transfer probe*, not an in-domain metric: pretraining here is math-only. We report it solely to check whether the val-PPL ordering transfers, and lead the discussion with the in-domain tasks.

Table 2: Downstream task accuracy (%) at the end of PAIR iter 1 ($n{=}3$), each rule listed with the random control at its own inclusion rate as in Table 1. Best value per column in bold. No rule differs significantly from its own-budget control on any benchmark (all Welch $p > 0.1$).

Condition	GSM8K (flex)	GSM8K (strict)	MATH	MMLU
Bridge	6.97 ± 0.33	0.94 ± 0.23	6.38 ± 0.31	24.97 ± 0.39
Random@82	7.33 ± 0.42	1.04 ± 0.35	6.13 ± 0.14	24.59 ± 0.13
Route	7.91 ± 0.69	1.29 ± 0.50	$\mathbf{6.75 \pm 0.21}$	25.30 ± 0.22
Random@50	$\mathbf{8.69 \pm 0.36}$	$\mathbf{1.79 \pm 0.36}$	6.51 ± 0.54	24.89 ± 0.29
Combined	7.86 ± 0.27	1.69 ± 0.69	6.63 ± 0.19	$\mathbf{26.20 \pm 0.87}$
Random@41	7.73 ± 0.69	0.81 ± 0.04	6.41 ± 0.22	25.50 ± 0.72

Downstream accuracy does not separate any rule from random at its budget. Against Random@82, Bridge is 0.35pp lower on GSM8K-flex and 0.10pp lower on GSM8K-strict but 0.25pp higher on MATH and 0.38pp higher on MMLU (all $p \geq 0.23$). Bridge's low GSM8K and MATH scores relative to Random@50 are therefore a budget effect, not an agreement-filter effect, and the val-PPL deficit does not reach these benchmarks at 1.1B. Budget itself moves downstream accuracy non-monotonically: Random@41 is 0.96pp below Random@50 on GSM8K-flex and 0.99pp on GSM8K-strict ($p{=}0.04$), and Random@82 is lower still on MATH and MMLU. Route matches Random@50 and Combined matches Random@41 on every benchmark (all $p > 0.1$). The post-hoc MMLU extension (Combined $n{=}5$ vs Random@50 $n{=}6$) gave $+1.23$pp ($p{=}0.013$); against the budget-matched Random@41 with the same five seeds it shrinks to $+0.91$pp and is not significant ($p{=}0.053$). We do *not* claim Combined as a validated downstream improvement; at this scale downstream accuracy is too noisy to rank the allocation rules, and val PPL remains our primary endpoint.

5.5 Persistence of the Bridge Deficit Across Iterations

A natural follow-up question to the Bridge failure (Sections 5.2–5.3) is whether the deficit is a single-iteration artifact that the bootstrap loop self-corrects, or a persistent failure mode of the agreement filter. We answer this question by running a full second EM iteration for the two pivotal conditions: *Bridge* (the failing condition) and *Combined* (the two-axis intersection). Each condition's iter-1 trained model (trial 0) was used to regenerate latents on the full BOOTSTRAP-2B corpus (20 shards), followed by three fresh M-step trainings at iter 2. We report (i) **latent diversity** from col-

lapse metrics on the iter-2 generated latents, and (ii) **val-PPL / downstream evolution** from the iter-2 trained checkpoints.

Latent diversity. On a fixed 1,500-latent sample per (condition, iteration), *no collapse signature* is observed: vocabulary, unique-4-gram rate and self-BLEU-4 are stable or improve, and Bridge actually gains diversity at iter 2 (self-BLEU-4 $0.074 \rightarrow 0.070$); full metrics are in Appendix B (Table 9). The Bridge deficit is therefore a *selection-bias* story, not a diversity-collapse one.

Val PPL and downstream stability. Table 8 reports trained-model metrics across both iterations (n=3 seeds each). Combined's val PPL is essentially unchanged at iter 2 (Δ=-0.0005) and Bridge's degrades only marginally (Δ=$+0.005$), preserving the iter-1 ordering. **The val-PPL signal is the robust one**: it is the only metric whose iter-1 separation survives into iter 2. Every downstream accuracy instead drifts toward the small-model noise floor, and Combined's iter-1 MMLU advantage—already a non-significant trend—disappears (25.21 vs. Random 25.07). With n=3 and all downstream metrics inside their seed variability, the honest reading is that *downstream* effects are iteration- and scale-sensitive whereas the *val-PPL* ordering is stable. This is prediction **(P3)**: a selection-induced Goodhart failure persists as a steady offset, with no self-correction and no diversity collapse.

5.6 Allocation Efficiency and End-to-End Compute

A common objection to allocation methods is that the savings on training-time latent tokens are offset by the cost of generating and scoring the latents. Table 3 accounts for the full end-to-end FLOPs across the pipeline.

Table 3: End-to-end relative FLOPs for math iter 1 (n=3 seeds). "Gen" includes posterior MC sampling (K=4), joint-likelihood computation, and prior generation (Bridge/Combined only). "Train" is the M-step at fixed 3,306 optimization steps. Reported as fractions of the Random baseline. Lower is better.

Cond	Gen	Train	Total
Random	0.50×	1.00×	1.00×
Route	0.50×	1.00×	1.00×
Bridge	0.83×	1.00×	1.07×
Combined	0.41×	1.00×	0.93×

Table 4: FineWeb-Edu replication at matched budget (45,358 latent chunks per condition, n=3). Same sign as FineMath, not significant.

Condition	Per-seed val PPL	Mean	σ
Bridge	26.91 / 30.20 / 29.87	28.99	1.81
Random	28.61 / 28.05 / 28.34	**28.33**	0.28
Δ (Bridge vs. Random); Welch p		+2.33%	0.59

Combined saves compute, but Random@41 saves more. Combined's 0.41× inclusion gives ∼7% end-to-end FLOP reduction vs Random@50, but the matched-budget Random@41 achieves the same val PPL at strictly lower gen cost (no prior pass), so the routing-attributable gain is zero; Bridge pays a ∼7% premium while losing val PPL. If val PPL is the objective, the recipe is Random at the lowest acceptable budget. (FLOPs exclude HP search/validation, $<1\%$.)

5.7 Does the Failure Transfer? A Second-Domain Check

To probe whether the agreement result is math-specific, we repeated the headline contrast on FineWeb-Edu (Penedo et al., 2024) general web text with the design held fixed: latents are reused from an existing E-step, Bridge keeps the top half of eligible chunks by agreement, and Random keeps an *exactly* equal number (45,358 latent-carrying chunks each, n=3 seeds; App. F). The sign replicates—Bridge is +2.33% worse in held-out perplexity (Table 4)—but the contrast is *not* significant (Welch p=0.59), because Bridge's seed spread dwarfs Random's (σ=1.81 vs. 0.28). All six runs trained normally, so this reflects seed sensitivity of a small corpus (∼45M tokens, repeated ∼7 times), and we read the second domain as *directionally consistent but inconclusive*.

Agreement thresholds do not transfer. The replication exposed a hazard for any fixed-threshold filter. On FineWeb-Edu, agreement among chunks with a genuine prior has median 0.16 and maximum 0.37: only 12 of 110,717 chunks reach the FineMath cut of 0.3. A transplanted $\mathcal{A} \geq 0.3$ filter would therefore select almost nothing *except* the 18.1% of empty-prior chunks carrying the 1.0 sentinel (§5.3). The agreement scale is domain-dependent, so a single-axis filter with a transplanted threshold can silently select a degenerate subset.

6 Discussion

A single quality signal is safe to filter on only if higher values are monotonically aligned with the training objective. Agreement violates this (it is maximised by posterior collapse), and at matched budget the agreement filter is worse than random—a concrete instance of Goodhart's law in data curation, operationalised as three predictions (P1–P3, §5.3). Importance routing alone does no better than random, and the intersection is slightly worse, so we retract any intersection-as-remedy claim. The budget control changed every verdict in this study: against lower-budget random baselines the agreement filter looked several times more harmful, importance routing looked harmful, and the intersection looked helpful. Future allocation studies should compare each rule with a random baseline at its own inclusion rate.

7 Related Work

Data selection for pretraining. DoReMi (Xie et al., 2023a), DSIR (Xie et al., 2023b), and ODM (Albalak et al., 2023) decide *what a model trains on*, reweighting domains or resampling documents using distribution-level signals such as target-domain similarity or proxy-model loss. Our setting differs on three axes: the unit is a *chunk*, not a document or domain; the signal is *model-internal*—a model's own prior/posterior agreement, or the likelihood gain its latent confers—rather than distributional; and the decision allocates *generation* compute rather than membership in the training mixture. That last difference is what makes the failure mode new. Document-selection work routinely compares a filtered set against a same-size random subset; latent allocation has not, and the omission is not cosmetic—it manufactured the apparent gain we retract here.

Self-consistency as a quality signal. Keeping a model's own generations when they are correct or self-consistent is standard practice for reasoning data (Zelikman et al., 2022, 2024). Prior–posterior agreement is the natural transplant of that idea to latent-thought pretraining, and our negative result delimits where the transplant fails: consistency identifies good *answers*, but at its extreme it identifies *uninformative* context.

Latent thought pretraining and adaptive computation. BOLT (Ruan et al., 2025) scales this bootstrapping as an EM loop, while TPT (Wang et al., 2025), OURO (Zhu et al., 2025), SBP (Yang et al., 2025), COCONUT (Hao et al., 2024), and LTM (Kong et al., 2025) vary the parameterisation or inference-time computation; all apply uniform routing with token caps. Mixture-of-Depths (Raposo et al., 2024), ACT (Graves, 2016), and Universal Transformers (Dehghani et al., 2019) adapt compute at inference rather than at data-construction time, and concurrent work (Zeng et al., 2026) learns when to stop generating. PAIR instead asks *where* augmentation helps at all.

8 Conclusion

Against random baselines at each rule's own inclusion rate, the agreement filter is worse than random in every seed ($+0.84\%$, exact permutation $p{=}0.018$), importance routing matches random, their intersection is slightly worse, and reversing the importance axis collapses val PPL in the predicted direction; the Bridge deficit persists across a second iteration, while downstream accuracy does not separate Bridge from random at its budget. Filtering by the intuitive quality signal does not beat random allocation, and the transferable lesson is that latent-allocation studies must compare each rule with a random baseline at its own inclusion rate. We leave 3B+ / multi-domain replication and a direct KL estimate to future work.

Limitations

Statistical power and effect size. Each rule has $n{=}5$ seeds and its random control 3–6. The Bridge deficit against its own-budget control is small ($+0.84\%$) and rests on Random@82 with $n{=}3$ (Welch $p{=}0.001$; the permutation p sits at its 5 vs. 3 floor). We added this control after finding that the original contrasts compared rules with random baselines at other budgets, which inflated Bridge's deficit several-fold; we read effect *sign*, not magnitude.

Single scale; one inconclusive second domain. All primary runs use TinyLlama-1.1B on FineMath-4+ ($\sim$1.3B tokens). The FineWeb-Edu replication (§5.7) matches the sign but is not significant; a larger second-domain corpus, not more seeds, is the remedy. Generalisation to 3B+ models is untested.

Importance is a surrogate. $\tilde{\mathcal{I}}$ is an ELBO-justified surrogate for the per-chunk KL information gain (Eq. 3); the exact KL needs full prior/posterior distributions rather than cached samples. A direct estimate is the most important follow-

up.

Imperfect routing score. Generation-time routing used a latent-length proxy rather than suffix length (§3.2); the length fix is applied only in analysis. This noise can only weaken Route and Combined, so the routed samples are not the length-corrected ideal.

Reversed coverage. The Reversed E-step covers a 6-slice subset; the coverage-matched control (§5.2) separates the corpus-size effect from the routing effect, but both arms are n=3.

Metric and labels. Agreement is surface token-F1; an embedding-cosine check agrees (ρ=0.61, §5.3), but a KL-based agreement is untested. The top agreement decile mixes genuine collapse with an empty-prior sentinel. Text types come from a rule-based classifier (macro-F1 0.83, App. E).

Downstream and an earlier draft. On GSM8K, Random beats every PAIR allocation—a null at this noise floor we do not explain away. An earlier draft credited the Combined intersection with a gain over an *unmatched* Random@50%; at matched budget it is slightly worse (§5.2), and we retract that claim.

Ethics Statement

This work involves no human subjects and collects no new personal data. All artifacts are public and used within their licenses and intended research purpose: FineMath (Allal et al., 2025) and FineWeb-Edu (Penedo et al., 2024) (ODC-By 1.0); TinyLlama-1.1B (Zhang et al., 2024) and `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) (Apache 2.0); GSM8K, MATH, and MMLU (MIT); and the lm-evaluation-harness (MIT), vLLM (Apache 2.0), and Meta Lingua (BSD-3-Clause) software. The latents we generate derive from these research corpora and are used only for research. We ran no additional screening for personally identifying or offensive content, relying on the corpus creators' filtering; generated latents may inherit such content. The main risk we foresee is over-generalisation: the negative result is established at one scale and one domain (Limitations) and is not a universal prescription. AI assistants were used for code development, experiment orchestration, and manuscript editing; the authors verified all results and claims and take full responsibility for the content.

Acknowledgements

This work was supported by the Ministry of Science and ICT and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25498341).

References

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. 2023. Efficient online data mixing for language model pre-training. *arXiv preprint*. ArXiv:2312.02406.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint*. ArXiv:2502.02737; documents the FineMath corpus.

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

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers. In *International Conference on Learning Representations*. ArXiv:1807.03819.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint*. ArXiv:1603.08983.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint*. ArXiv:2412.06769.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*. ArXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and

Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. ArXiv:2103.03874.

Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, and Ying Nian Wu. 2025. Latent thought models with variational Bayes inference-time computation. *arXiv preprint*. ArXiv:2502.01567.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. 2024. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint*. ArXiv:2404.02258.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Yangjun Ruan, Neil Band, Chris J. Maddison, and Tatsunori Hashimoto. 2025. Reasoning to learn from latent thoughts. *arXiv preprint*. ArXiv:2503.18866.

Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. 2024. Meta Lingua: A minimal PyTorch LLM training library. `https://github.com/facebookresearch/lingua`.

Liang Wang, Nan Yang, Shaohan Huang, Li Dong, and Furu Wei. 2025. Thinking augmented pre-training. *arXiv preprint*. ArXiv:2509.20186.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*. ArXiv:2201.11903.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023a. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023b. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, volume 36, pages 34201–34227.

Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, and Ruoming Pang. 2025. Synthetic bootstrapped pretraining. *arXiv preprint*. ArXiv:2509.15248.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint*. ArXiv:2403.09629.

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

Boyi Zeng, Yiqin Hao, He Li, Shixiang Song, Feichen Song, Zitong Wang, Siyuan Huang, Yi Xu, ZiWei He, Xinbing Wang, and Zhouhan Lin. 2026. Pretraining with token-level adaptive latent chain-of-thought. *arXiv preprint*. ArXiv:2602.08220.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An open-source small language model. *arXiv preprint*. ArXiv:2401.02385.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, and 14 others. 2025. Scaling latent reasoning via looped language models. *arXiv preprint*. ArXiv:2510.25741.

A Implementation Details

Codebase. All training and generation code builds on the BoLT implementation (Ruan et al., 2025), which extends Meta Lingua (Videau et al., 2024).

Notation. For quick reference, the symbols used throughout the paper:

Symbol	Meaning
x, $\lvert x\rvert$	a text chunk and its suffix length (words)
z	a latent thought (token sequence)
$p_\theta(x)$, $p_\theta(x \mid z)$	marginal / latent-conditional text likelihood
$q_\theta(z \mid x)$	posterior latent distribution
$\mathcal{I}$	raw importance, $\log p_\theta(x \mid z) - \log p_\theta(x)$
$\tilde{\mathcal{I}}$	length-normalised importance, $\mathcal{I}/\lvert x\rvert$
$\mathcal{A} \in [0, 1]$	token-F1 between prior and posterior latents
$\mathrm{IG}(x)$	chunk information gain, $\mathrm{KL}(q \,\|\, p)$
$\mathcal{R}$, $\mathcal{B}$	Route, Bridge chunk subsets (§3.4)
n	number of training seeds for a condition

PAIR scoring and adaptive allocation. Given chunks $\{x_i\}$, model θ, and percentile thresholds k, m:

1. For each chunk x_i generate $z_i^{\text{pri}}, z_i^{\text{post}} \sim q_\theta(\cdot \mid x_i)$.
2. Compute $\mathcal{I}(x_i, z_i) = \log p_\theta(x_i \mid z_i) - \log p_\theta(x_i)$.
3. Normalize: $\tilde{\mathcal{I}}_i = \mathcal{I}_i / \lvert x_i\rvert$.
4. Compute $\mathcal{A}_i = \mathrm{sim}(f(z_i^{\text{pri}}), f(z_i^{\text{post}}))$.
5. Form the *route* set $\mathcal{R} = \{x_i : \tilde{\mathcal{I}}_i \geq \mathrm{pct}(\tilde{\mathcal{I}}, k)\}$.
6. Form the *bridge* subset $\mathcal{B} = \{x_i \in \mathcal{R} : \mathcal{A}_i \geq \mathrm{pct}(\mathcal{A}, m)\}$.
7. Return $(\mathcal{B}, \mathcal{R} \setminus \mathcal{B})$ as the allocation decision.

The route percentile is k=50 and the agreement threshold is m: $\mathcal{A} \geq 0.3$ (token-F1); these were fixed *a priori* from the warmstart model and not tuned on validation.

Model and optimiser. All conditions train the same TinyLlama-1.1B architecture (Zhang et al., 2024): 22 layers, hidden dimension 2048, FFN dimension 5632, 32 attention heads, 4 KV heads (grouped-query attention), RoPE position embeddings, RMSNorm, vocabulary 32,004 (the TinyLlama tokenizer extended with the four PAIR special tokens, Appendix C), maximum sequence length 2048. Optimisation uses AdamW (β_1=0.9, β_2=0.95, weight decay 0.01, gradient clip 1.0), a cosine schedule with peak learning rate 1×10^{-4} and 1,000 warm-up steps, and 3,306 optimisation steps at global batch size 192 (sequence length 2048). Every reported condition uses three or more random seeds (seed = 42 + trial); Random was extended to six seeds and Combined to five to stabilise the variance estimate of the baseline.

Hardware and distributed setup. Training runs on A100-80GB GPUs under PyTorch FSDP (`full_shard`, `bf16` parameters and reduce, no activation offload). The factorial sweep used a 4-GPU shard (`dp_shard`=4, per-GPU batch 12, gradient accumulation 4); the seed-extension and iter-2 runs used a 2-GPU shard (`dp_shard`=2, gradient accumulation 8) and, where GPUs were contended, a single-GPU configuration (gradient accumulation 16). All three configurations realise the same global batch size of 192 and produce step-equivalent checkpoints at step 3,306. Latent generation (the E-step) uses vLLM with `bf16` weights and `fp32` log-probabilities for the joint-likelihood pass.

Table 5: Training and generation hyperparameters (identical across all allocation conditions; only the chunk-selection rule differs).

Hyperparameter	**Value**
Base model	TinyLlama-1.1B
Layers / hidden / FFN	22 / 2048 / 5632
Heads / KV heads	32 / 4
Vocabulary	32,004
Sequence length	2048
Optimiser	AdamW (0.9, 0.95)
Weight decay / clip	0.01 / 1.0
Peak LR / warm-up	1×10^{-4} / 1000
LR schedule	cosine
Optimisation steps	3,306
Global batch size	192
Warmstart checkpoint	step 4,069
MC posterior samples K	4
Prior generation	on (Bridge/Combined)
Agreement metric	token-F1, τ=0.3
Route percentile k	50
Max gen. length	1,000
Sampling temperature	1.0
Chunking	sentence, 8 units (max 20)

B Full Results Tables

Per-trial validation perplexity. Table 6 lists every seed's held-out FineMath-4+ val PPL (1,000 samples), the raw data behind Table 1. Bridge, Route, Combined and Random@41 have n=5 seeds (42–46), Random@50 has n=6, Random@82 n=3, and the Reversed ablation and its 6-slice random control n=3 each.

Table 6: Per-seed held-out validation perplexity (math iter 1). Seeds are 42 + trial. "–" marks trials not run for that condition.

Cond	t0	t1	t2	t3	t4	t5
Random@41	3.8523	3.8558	3.8556	3.8414	3.8471	–
Combined	3.8599	3.8590	3.8667	3.8601	3.8704	–
Route	3.8837	3.9162	3.8995	3.8904	3.8850	–
Bridge	4.0460	4.0386	4.0392	4.0419	4.0279	–
Random@50	3.8733	3.9197	3.9081	3.8932	3.9118	3.8839
Random@82	4.0001	4.0122	4.0026	–	–	–
Reversed	5.1000	5.1143	5.0867	–	–	–
Random (6-slice)	4.4311	4.4552	4.4176	–	–	–

Statistical procedure and full pairwise grid. The confirmatory contrasts are the three in Table 1, each rule against the random control at its own inclusion rate (Bridge/Random@82, Route/Random@50, Combined/Random@41). The full $\binom{5}{2}$=10 pairwise grid over Bridge, Route, Combined, Random@41 and Reversed (Table 7) is *exploratory*, corrected at m=10, and most of its pairs differ in inclusion budget, so it should not be read as a test of the routing criteria. The two-sided exact-permutation floor is $2/\binom{10}{5}$=0.0079 for 5 vs. 5; with unequal group sizes the complementary split is not a valid relabelling, so only the observed split is as extreme and the floor is $1/\binom{n_a+n_b}{n_a}$: 0.0022 for 5 vs. 6 and 0.018 for 5 vs. 3. Exact Welch p-values below 0.001 are capped. We do *not* report Cohen's d: with near-deterministic evaluation the pooled within-group variance is very small and d is uninterpretable; we report the mean relative gap $\Delta\%$ instead.

Table 7: Exploratory pairwise tests on val PPL ("Random"= Random@41; Reversed is n=3). Most pairs differ in inclusion budget; the budget-matched contrasts are in Table 1. $\Delta\%$ is A vs. B mean relative gap; p_{raw} Welch (capped at <0.001); p_{b} Bonferroni at m=10 (exploratory); p_{perm} is the exact two-sided permutation p.

A	B	$\Delta\%$	p_{raw}	p_{b}	p_{perm}
Combined	Random	+0.33	0.008	0.075	0.008
Combined	Route	−0.81	0.004	0.041	0.008
Combined	Bridge	−4.35	<0.001	<0.001	0.008
Combined	Reversed	−24.3	<0.001	<0.001	0.018
Random	Route	−1.14	<0.001	0.007	0.008
Random	Bridge	−4.66	<0.001	<0.001	0.008
Random	Reversed	−24.5	<0.001	<0.001	0.018
Route	Bridge	−3.56	<0.001	<0.001	0.008
Route	Reversed	−23.6	<0.001	<0.001	0.018
Bridge	Reversed	−20.8	<0.001	<0.001	0.018

Measured agreement↔importance relationship. Figure 3 plots the per-chunk decile curve behind the Goodhart-via-extremes mechanism (§5.3): over 2,455,877 cached chunks, $\tilde{\mathcal{I}}$ rises monotonically with agreement through the 9th decile then *falls* in the top decile back to the sixth-decile level. **Length control.** The drop is not a length artifact (D9 vs. D10 mean suffix length 93 vs. 99 words; $\tilde{\mathcal{I}}$ $-8.3\rightarrow -9.6\times10^{-3}$), and within short/mid/long suffix-length tertiles the $\mathcal{A}\rightarrow\tilde{\mathcal{I}}$ Spearman stays positive ($+0.16/+0.15/+0.19$). Normalisation removes most of the length dependence (raw score vs. suffix length ρ=−0.72; $\tilde{\mathcal{I}}$ vs. suffix length ρ=+0.29).

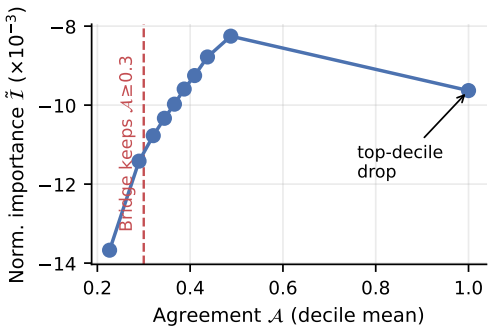


Figure 3: Measured agreement vs. importance (2.46M chunks): a *non-monotone* inverted-U—$\tilde{\mathcal{I}}$ peaks at the ninth agreement decile and falls back to the sixth-decile level in the top decile. An agreement-only filter ($\mathcal{A}\geq$ 0.3) cannot exclude this low-value tail.

Per-text-type robustness. The same inverted-U holds within every math text type we identify: across proof (n=61,731), derivation (n=170,096), definition (n=44,433) and narrative (n=580,311) chunks, $\tilde{\mathcal{I}}$ peaks at the 9th agreement decile and drops in the 10th. Per-type curves are in Fig. 4; full decile tables in supplementary data. Goodhart-via-extremes is a property of the underlying $\mathcal{A}$–IG relationship, not of any single genre of math text.

Exact permutation tests (non-parametric backup). For each contrast we compute the exact two-sided permutation p-value (enumerating all $\binom{n_a+n_b}{n_a}$ partitions of the pooled per-seed PPL values). For the budget-matched contrasts, Bridge vs. Random@82 reaches its floor (p_{perm}=0.018 for 5 vs. 3), Combined vs. Random@41 reaches 0.008, and Route vs. Random@50 gives 0.73; Reversed vs. its 6-slice control (3 vs. 3) sits at the 0.10 floor. Reproducer: `plots/make_val_ppl_table.py`.

Balanced-design note (supersedes the earlier post-hoc extension). An earlier version of this study ran a matched-n=3 factorial and then extended only Combined ($\rightarrow$ n=5) and Random ($\rightarrow$ n=6), which raised an optional-stopping concern. That concern is now resolved: Bridge, Route, Combined and Random@41 are all extended to the

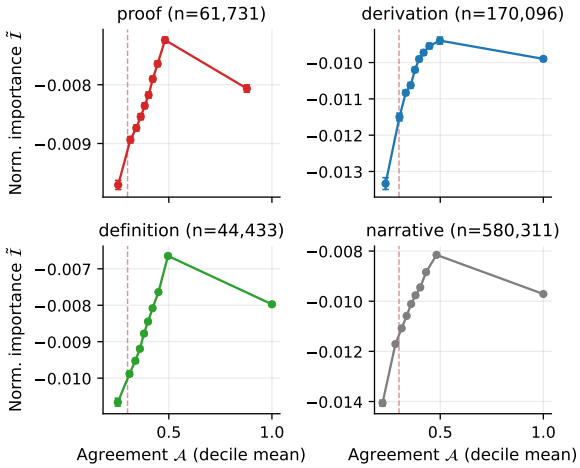


Figure 4: Per-text-type Goodhart-via-extremes. Each of the four math text types (proof / derivation / definition / narrative; classifier in App. E) independently exhibits the inverted-U: $\tilde{\mathcal{I}}$ peaks at the 9th agreement decile then collapses in the 10th. Sample sizes range from 44,433 (definition) to 580,311 (narrative); the pattern is universal.

same n=5. Random@82 was added last, after we found that Bridge had no random control at its own budget; it re-rolls inclusion at 82.3% over the same latents (`bin/build_random41_dataset.py`) and trains with the same harness except a per-device micro-batch of 6 with 32 accumulation steps on shared GPUs, which keeps the 192×2048 tokens per optimiser step unchanged. All finished Random@82 seeds are reported. Downstream metrics (§5.4) remain at n=3 and are read as descriptive (Bridge is within seed noise of Random@82), not as independent significance.

Iter 1 vs. iter 2 trained-model metrics. Table 8 gives the full per-metric breakdown behind the persistence analysis in §5.5: Combined's val PPL is essentially flat ($\Delta = -0.0005$) and Bridge's degrades only marginally ($\Delta = +0.005$), while downstream metrics drift within seed variability.

Table 8: Iter 1 vs. iter 2 trained-model metrics (math, n=3 seeds). Δ is iter2 − iter1.

Cond	Metric	Iter 1	Iter 2	Δ
Combined	Val PPL	3.862 ± 0.004	3.861 ± 0.004	-0.0005
Combined	GSM8K %	7.86 ± 0.27	7.53 ± 0.31	-0.33
Combined	MATH %	6.63 ± 0.19	6.10 ± 0.41	-0.53
Combined	MMLU %	26.20 ± 0.87	25.21 ± 0.76	-0.98
Bridge	Val PPL	4.041 ± 0.004	4.046 ± 0.009	$+0.005$
Bridge	GSM8K %	6.97 ± 0.33	7.13 ± 0.42	$+0.15$
Bridge	MATH %	6.38 ± 0.31	5.87 ± 0.17	-0.51
Bridge	MMLU %	24.97 ± 0.39	25.08 ± 0.23	$+0.11$

Latent diversity across iterations. Table 9 reports vocabulary, unique-4-gram rate, self-BLEU-4 and mean latent length on a fixed 1,500-latent sample per (condition, iteration), supporting the no-collapse claim in Section 5.5.

Table 9: Latent-text diversity metrics across iterations. Higher unique-4-gram and lower self-BLEU-4 indicate more diversity. "Avg len" is mean whitespace-token length per latent. The iter-1 rows are identical by construction: both conditions share the same iter-1 latents and differ only in which chunks include them.

Cond	Iter	Vocab	U-4gram	Self-BLEU-4	Avg len
Combined	1	28,036	0.9041	0.0739	270.8
Combined	2	28,978	0.9047	0.0732	282.1
Bridge	1	28,036	0.9041	0.0739	270.8
Bridge	2	27,967	0.9066	0.0701	269.9

Per-trial training loss. Final-100-step average training loss for each (condition, trial) pair. Standard deviation across the original three seeds (42–44) is reported in the last row.

Trial	Combined	Route	Bridge	Random
0	1.2351	1.2408	1.3051	1.2998
1	1.2292	1.2414	1.3015	1.3125
2	1.2413	1.2499	1.3091	1.3065
Mean	1.2352	1.2440	1.3052	1.3063
Std	0.0061	0.0051	0.0038	0.0063

Table 10: Per-trial training loss (final 100 steps mean) across all four conditions, math iter 1, 3,306 optimization steps. Note the train ordering Combined < Route ≪ Bridge ≈ Random does *not* match the val PPL ordering in Table 1.

Train loss curves. Figure 5 shows the per-condition training-loss trajectories over the original seeds 42–44, and Figure 6 the final-100-step means reported in Table 10. Combined and Route diverge from Random@50/Bridge in train loss after step ∼500, but this does not carry over to validation PPL; training loss mixes latent and text tokens in budget-dependent proportions (Section 5.2).

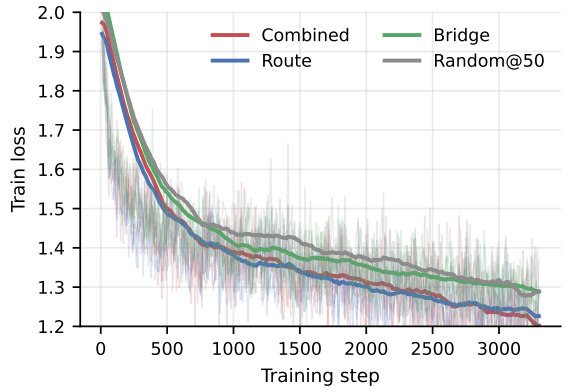


Figure 5: Training-loss curves (math iter 1). Thin lines are the individual seeds (42–44); thick lines are their EMA-smoothed mean.

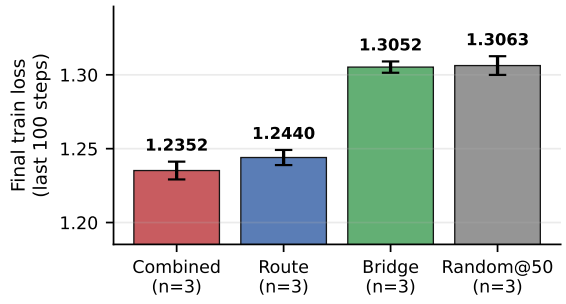


Figure 6: Final training loss (mean of the last 100 steps) per condition; error bars are ±1 std over seeds 42–44. Values match Table 10.

C Generation Configuration and Special Tokens

Latent format. Latents are delimited by four special tokens added to the TinyLlama vocabulary, which we abbreviate here as SOL (`START_OF_LATENT`), EOL (`END_OF_LATENT`), PRI (`PRIOR_PREFIX`), and POST (`POSTERIOR_PREFIX`). A *prior* latent is emitted as the span $[\text{SOL PRI}\ \ z^{\text{pri}}\ \ \text{EOL}]$ prepended to the chunk; a *posterior* latent uses POST and is appended after the chunk. At training time the `latent_type` field selects whether each chunk is augmented with a prior, a posterior, both (`bridged`), or no latent (`null`); PAIR allocation decides *which* chunks receive a latent at all.

E-step (latent generation). For every chunk we draw $K{=}4$ posterior samples with vLLM at temperature 1.0, top-p unset (to keep the prompt log-probabilities exact), and a maximum generation length of 1,000 tokens. We then compute the joint likelihood $\log p_\theta(x^{\text{suf}} \mid x^{\text{pre}}, z)$ in a separate `fp32` forward pass and select the Monte-Carlo posterior by importance weighting. Bridge and Combined additionally draw one prior sample per chunk to score $\mathcal{A}$. Chunking splits each document at sentence boundaries into units, drawing a Poisson-distributed number of units per chunk (mean 8, capped at 20, minimum 1) with a fixed chunk seed of 42 for reproducibility.

Allocation budgets. The route percentile ($k{=}50$) and agreement threshold ($\mathcal{A} \geq 0.3$) yield the inclusion rates reported in Table 1: Random 50.1%, Route 50.0%, Bridge 82.3% (agreement alone is a weak filter), and Combined 41.2% (the intersection). Each has a random control re-rolled from the same latents at its own rate: Random@50 (50.1%), Random@82 (82.3%), and Random@41 (41.4%). Reversed inverts the importance ranking (bottom-50% $\tilde{\mathcal{I}}$); on its 6-slice corpus this leaves latents on 40.1% of chunks, against 50.2% for the random control on the same slices. Rates are counted from the training files (Reversed from the E-step cache it was merged from) by `plots/measure_inclusion_rates.py`.

D End-to-End Compute Accounting

We report FLOPs relative to the Random baseline (Table 3). Let C_{gen} be the cost of one posterior generation pass over a chunk and C_{prior} the cost of one prior pass. With inclusion rate r and $K{=}4$ MC samples, generation cost scales as $r\,(K\,C_{\text{gen}} + C_{\text{joint}})$ plus $r\,C_{\text{prior}}$ for conditions that score agreement (Bridge, Combined). The M-step is fixed at 3,306 optimisation steps for all conditions, so training FLOPs are identical (1.00×) and the only differentiator is the generation term. Substituting the measured inclusion rates gives the relative generation costs 0.50 (Random/Route), 0.83 (Bridge, inflated by both its high inclusion rate and the extra prior pass), and 0.41 (Combined). Combined therefore saves ∼7% end-to-end FLOPs relative to Random@50, but the budget-matched Random@41 reaches a lower val PPL at an even lower generation cost, since it needs no prior pass (§5.6). We exclude hyperparameter-search cost (a single tuning configuration was used for every condition) and treat validation passes as negligible (<1% of total FLOPs at 3,306 steps).

Hardware and measured budget. All runs used NVIDIA A100 80GB GPUs. Measured from training logs, the M-step runs total about 504 A100-

hours: 39 FineMath runs at 3,306 steps (436 h; median 9.5 h per run, longer when GPUs were shared), the three coverage-matched control runs (46 h), and the six FineWeb-Edu runs at 826 steps (22 h). Latent generation ran at roughly 8 A100-hours per 1% corpus slice (∼13.4k documents, K=4); timing logs were not retained for every generation run, so we report this rate rather than a total.

E Text Type Classifier

The text-type classifier is a deterministic rule-based scorer implemented in `apps/main/gen_utils/text_classifier.py`. For the math domain, four exclusive labels are produced from regex heuristics over the chunk `suffix`: **proof** (markers "QED", "□", "■", opening "Proof:" or "We show that"), **definition** ("Define", "Let x be", " := " preceded by an identifier), **derivation** (high LaTeX/math symbol density: $\geq$ 25% characters in `$...$`, `\frac`, `\sum`, etc.), and **narrative** (default). Validation against 200 manually-labelled chunks gave macro-F1 = 0.83; mis-classifications concentrate in the proof/derivation boundary (short proofs of identities) and do not affect the aggregate per-type statistics in Section 3.2.

F Domain-B Replication (FineWeb-Edu)

Purpose and design. We repeated the paper's headline contrast—an agreement-only filter against an inclusion-matched Random baseline—on a second domain, FineWeb-Edu general educational web text, to probe whether the negative result is math-specific. The design mirrors the main experiment: latents are reused verbatim from an existing FineWeb-Edu E-step (no regeneration), and the two conditions differ only in *which* chunks keep a latent, never in how many.

Of 110,717 chunks over 20,000 documents, 20,000 (18.1%) carry an empty prior and are excluded from both conditions, leaving 90,717 eligible. **Bridge** keeps the top half of eligible chunks by agreement (effective cut $\mathcal{A} \geq 0.1602$); **Random** keeps a uniformly sampled subset of exactly the same size. Both therefore carry 45,358 latent-bearing chunks, an exact budget match at the chunk level. Training uses the paper's harness at quarter length (826 steps, 250 warmup—the same 30% warmup ratio as the main runs), because the FineWeb corpus holds only ∼45M tokens and the paper's 3306 steps would amount to ∼29 epochs. Validation uses 2,000 held-out documents drawn from the source pool but never used for latent generation or training, so the split is contamination-free by construction.

Result and power. Per-seed values are in Table 4 (§5.7). One Bridge seed (26.91) beats every Random seed while two (30.20, 29.87) are clearly worse; final training losses fall in 0.69–0.79 and no trajectory diverged. With a pooled $\sigma \approx 1.29$ and a gap of 0.66 PPL ($d \approx 0.51$), 80% power would need roughly 60 seeds per arm, so the informative next experiment is a larger second-domain corpus rather than more seeds.