COUPLED AUTOREGRESSIVE DECODING: MARGINAL CONTRACTS AND SUCCESS GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating several answers gives a language model multiple chances to solve a problem. The benefit depends on whether those answers succeed together or cover one another's failures. Coupled decoding changes this dependence while preserving each answer's distribution. We study when coordinating answers in this way increases the chance of obtaining at least one correct answer. We compare two requirements for matching the original model: preserving each complete answer's distribution, and additionally preserving its next-token probabilities after observing the past tokens of every answer in the batch. The additional restriction limits coordination and can reduce the highest achievable success probability. Neither requirement alone guarantees improvement. To understand which sampling methods help, we use an arithmetic decoder that generates a complete answer from a single number between zero and one. Each answer corresponds to an interval of numbers that generate it, so correct answers occupy a collection of intervals. Dividing the original interval into equal pieces and drawing one number uniformly and independently from each piece provides a guarantee: with a common decoder and correctness rule, the chance of obtaining a correct answer is at least as high as with independent uniform draws from the whole interval. This guarantee does not generally hold for equally spaced points moved together by a random amount, or for sampling anew within equal pieces at each token. Both alternatives can help or hurt. Reordering token intervals can even reverse the effect of equally spaced sampling without changing individual answer probabilities. Experiments find no supported gain for the selected couplings on open-ended math, but a gain from equally spaced sampling on multiple-choice Countdown. A separate Countdown study finds a moderate association between estimated success geometry and independently measured gains.

1 INTRODUCTION

When a language model generates several answers, the batch succeeds if at least one answer is correct. For K answers, this probability is called pass@K. Their dependence can matter as much as individual accuracy. Consider two answers, each correct half the time. If their successes always coincide, the pair succeeds only half the time; if their successes are complementary, it always succeeds. Coupled decoding exploits this possibility by changing dependence across answers while preserving each answer's distribution.

Existing methods, including Arithmetic Sampling and quasi-random decoding, use structured randomness to generate such batches (Vilnis et al., 2023; Amballa et al., 2025; Li et al., 2026). Spreading samples out aims to use the generation budget better, but need not spread them across correct answers. A construction can preserve every individual output distribution yet make the batch less likely to solve the problem. We ask which constructions offer protection against this failure, and what that protection requires.

We distinguish two requirements for preserving the base model. The weak requirement preserves each complete answer's distribution. The strong requirement also preserves its next-token distribution given the past tokens of every answer in the batch. Other answers can reveal information about an answer's future without changing its individual distribution, so these requirements differ. The stronger requirement can reduce the best attainable batch success.

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.

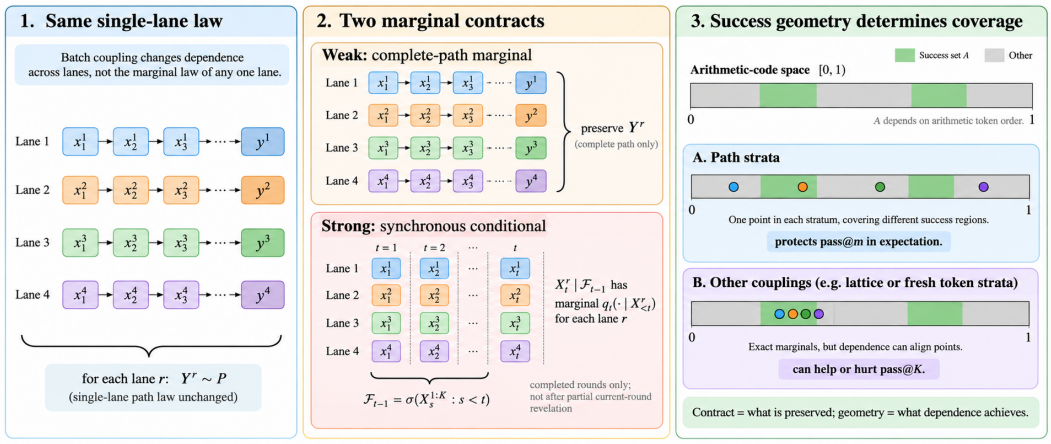

Figure 1: Marginal contracts and success geometry, illustrated for a common base decoder. Both contracts preserve individual answer distributions; the strong contract also preserves next-token conditionals given completed joint rounds, excluding current-round tokens and private randomness. Independent path strata draw independently within equal subintervals, then uniformly assign points to lanes (individual generations). For any fixed $m \leq K$ lanes, pass@m is at least IID's in expectation over sampling. Lattices and fresh token strata can help or hurt. Points are schematic, not literal lattice locations. Arithmetic token order rearranges success intervals without changing path probabilities.

We next identify constructions that protect batch success. An arithmetic decoder maps points in an interval to complete answers; correct answers occupy regions of that interval. With a shared decoder and success rule, independent draws within equal subintervals cannot reduce batch success relative to independent and identically distributed (IID) sampling. This guarantee averages over sampler randomness and covers fixed batch subsets.

Equally spaced points moved by a shared random shift form a lattice. It can increase or decrease batch success. Redrawing points within equal subintervals at each token also lacks this protection. Even changing the order of token intervals in the arithmetic decoder can turn a lattice's improvement over independent sampling into a loss, while leaving every answer's probability unchanged. Figure 1 illustrates these distinctions; Table 1 summarizes the guarantees.

Exact comparisons on two-step decoders exhibit both gains and losses. The selected couplings show no supported improvement on the tested open-ended math problems. On compact multiple-choice Countdown, the lattice improves batch success. A separate Countdown study finds a moderate association between estimated success geometry and independently measured gains. These results distinguish guaranteed protection from opportunities for improvement.

Related work. We connect marginal-preserving decoding, transport constraints, and classical probability inequalities. Arithmetic Sampling maps structured uniform points through arithmetic codes while preserving marginals (Vilnis et al., 2023); quasi-random multi-sample inference applies this idea to reasoning and translation (Amballa et al., 2025). QuasiMoTTo studies lattice, path-stratified, and token-level constructions, union bounds for batch success, and estimators of pass@K for correlated samples (Li et al., 2026). Our protection theorem and counterexamples distinguish the guarantees of these constructions. Our Countdown diagnostic uses QuasiMoTTo's lattice subsampling structure with new data and another model. The contracts use conditional-law constraints from causal, adapted, and multicausal transport (Backhoff et al., 2017; Eckstein & Pammer, 2024; Acciaio et al., 2025). The entropy and success-count results specialize information projection (Csiszár, 1975) and Bernoulli-sum ordering (Hoeffding, 1956) to decoding. Corvelo Benz et al. (2026) couple different models to reduce evaluation variance; we study same-decoder batch success. Diverse beam search, sampling without replacement, and determinantal beam methods impose different constraints on individual output distributions (Vijayakumar et al., 2018; Kool et al., 2019; Meister et al., 2021).

Table 1: Guarantees of the evaluated constructions. Path marginals concern each complete answer's distribution. Synchronous fidelity additionally preserves next-token distributions given the completed batch history. Pass protection compares expected success with IID for fixed lane subsets sharing a decoder, ordering, and success rule. "No" indicates a counterexample, not universal harm.

Construction	Path marginals	Synchronous fidelity	Pass protection
IID	Preserved	Preserved	Reference
Independent path strata	Preserved	Not generally	Yes
Random-shift lattice	Preserved	Not generally	No
Fresh token strata	Preserved	Preserved	No

2 Marginal contracts and their attainable utilities

Preserving complete-answer distributions and next-token distributions given the batch history impose different constraints. We classify constructions under both contracts and show that the stronger one can lower optimal success. We quantify this cost for general batch scores with a penalty for departure from independence.

2.1 Complete paths versus synchronous conditionals

The strong contract implies the weak one, but the converse fails: other answers' pasts can reveal an answer's future. We first specify the decoder and available history.

Each generation is a lane; its complete token sequence is a path. Fix K named lanes, a finite token alphabet, and a finite horizon T. Write X_t^r for lane r's token at round t, $Y^r = X_{1:T}^r$ for its path, and P_r for its distribution. Next-token conditionals $q_{r,t}(\cdot \mid X_{<t}^r)$ incorporate temperature and probability filters. Lanes stop at an end-of-sequence (EOS) token, an answer marker, or a length cap, and remain stopped.

Independent decoding has distribution $Q = \bigotimes_{r=1}^K P_r$; the P_r may differ. A coupling changes dependence while preserving each P_r. Write $\Pi(P_1, \dots, P_K)$ for all such joint distributions.

At each *synchronous joint round*, every active lane generates its next token. The history $\mathcal{F}_{t-1} = \sigma(X_s^{1:K} : s < t)$ contains all completed rounds but neither another lane's current-round token nor the sampler's private current-round randomness. The joint next-token distribution (joint kernel) is chosen using this history.

Definition 1 (Two marginal contracts)**.** *The weak contract fixes complete-path marginals:*

$$\mathfrak{S}(P_{1:K}) = \Pi(P_1, \dots, P_K). \tag{1}$$

The strong class $\mathfrak{C}(P_{1:K}) \subseteq \mathfrak{S}(P_{1:K})$ additionally requires, at every positive-probability completed batch prefix,

$$\Pi(X_t^r \in \cdot \mid \mathcal{F}_{t-1}) = q_{r,t}(\cdot \mid X_{<t}^r), \qquad r = 1, \dots, K. \tag{2}$$

Arithmetic decoding illustrates the distinction. It partitions the unit interval into token intervals of widths equal to their probabilities, then subdivides the selected interval for each subsequent token. A single uniform point therefore determines a complete answer with the base decoder's path distribution.

A two-step example makes the distinction concrete. Take two uniform arithmetic-code points separated by $1/2$. Let the first-token ordered probabilities be $(0.7, 0.3)$ and the second-token probabilities be $(0.8, 0.2)$. Each individual code point generates the base decoder's path distribution.

Now observe that both lanes selected the first token's 0.7 interval. This restricts the possible code points, changing the first lane's probability of selecting the next 0.8 interval to 0.65. Each path distribution is preserved, but the next-token conditional is not (Appendix A.2).

Conversely, composing joint kernels in $\Pi(q_{1,t}(\cdot \mid h^1), \dots, q_{K,t}(\cdot \mid h^K))$ at completed histories $h = (h^1, \dots, h^K)$ satisfies Eq. (2) and preserves path marginals by induction. Observing tokens within a round can change the remaining tokens' conditional distributions. Thus, arbitrary within-round scheduling or pausing is not covered.

Concrete constructions. Marginally uniform code points chosen once per path preserve path distributions. Fresh uniforms at each round can also preserve next-token conditionals.

Write $D_r(u)$ for the arithmetic decoder in lane r. We compare three code-point constructions:

Arithmetic-IID draws independent $U_r \sim \mathrm{Unif}[0,1)$.

Random-shift lattices use equally spaced points with wraparound: $U_r = (S + (r-1)/K) \bmod 1$, where $S \sim \mathrm{Unif}[0,1)$.

Independent path strata draw independently within equal subintervals: $U_j = (j - 1 + V_j)/K$ where the V_j are independent $\mathrm{Unif}[0,1)$ offsets.

The latter two constructions uniformly permute points over named lanes. All three use decreasing filtered-probability token order. Each assigned point is marginally uniform, which proves the weak contract.

This argument assumes exact arithmetic, continuous uniforms, and fixed next-token distributions that depend only on each lane's own prefix; Appendix B.1 explains the numerical limits.

Fresh token strata redraw their randomness at each joint round. They divide the unit interval equally among active lanes, draw independently within each subinterval, and uniformly assign the points to lanes. This randomization is independent of the completed history.

Each active lane receives a marginally uniform point. Inverse-CDF sampling maps it to a token using cumulative probabilities, with tokens ordered by *increasing* filtered probability. Draws remain private until the round is sampled, satisfying the strong contract; stopped lanes remain stopped. Appendix Table 5 summarizes the contract differences.

2.2 A STRICT COST OF THE STRONG CONTRACT

Preserving next-token distributions under the strong contract can reduce optimal batch success even for identical decoders. These sequential constraints, called causal, can prevent arranging the lanes so that successes never overlap. A two-step decoder demonstrates the loss.

Let lane r succeed with probability s_r. With only path marginals fixed, the largest possible probability of at least one success is $\min\{1, \sum_r s_r\}$. Appendix Proposition 4 gives both attainable bounds.

For two lanes sharing a two-step decoder, let the first token equal one with probability $3/4$. Terminal success probabilities are $1/20$ and $13/20$ after prefixes zero and one, respectively, so each lane succeeds with probability $1/2$. The weak contract permits disjoint successes; the strong contract also constrains outcomes after each joint first-token history.

Proposition 1 (Strict causal gap)**.** *There is a two-lane, two-step decoder with success marginal* $1/2$ *in each lane whose optimal weak pass@2 is* 1*, but whose optimal strong pass@2 is* $17/20$*.*

Preserving intermediate conditionals causes the loss, although both contracts preserve individual success probabilities. Appendix A.3 derives both optima and shows they are attainable.

2.3 ENTROPY-REGULARIZED STATIC AND ADAPTED TRANSPORT

Penalizing departure from independence gives an exact cost for the strong contract: a scaled divergence between optimal joint distributions. Static transport optimizes under the weak contract; adapted transport enforces the strong conditional constraints.

Let G be a bounded batch score; for batch success, it is one if any answer is correct and zero otherwise. The parameter $\varepsilon > 0$ controls the penalty for departure from independence, measured by Kullback–Leibler divergence (KL). Define

$$L(\Pi) = \mathbb{E}_\Pi G - \varepsilon \mathrm{KL}(\Pi \| Q), \qquad \mathrm{OPT}_{\mathcal{A}} = \max_{\Pi \in \mathcal{A}} L(\Pi), \quad \mathcal{A} \in \{\mathfrak{S}, \mathfrak{C}\}. \tag{3}$$

On the finite path space supported by Q, let Π_w^* and Π_c^* be the unique weak and strong optimizers. The identities derived in Appendix A.4 yield the following comparison.

Corollary 1 (Exact entropic price of causality)**.** *For the same utility and base marginals,*

$$\mathrm{OPT}_{\mathfrak{S}} - \mathrm{OPT}_{\mathfrak{C}} = \varepsilon \mathrm{KL}(\Pi_c^* \| \Pi_w^*). \tag{4}$$

Equivalently, Π_c^* *minimizes* $\mathrm{KL}(\Pi \| \Pi_w^*)$ *over* $\mathfrak{C}$.

The cost vanishes if the weak optimum already satisfies the strong contract. Otherwise, it is their directed KL divergence scaled by ε.

These are ideal optima: computing the adapted optimum requires future batch scores. The constructions above are not designed to attain them. Appendix A.4 gives the optimizer characterizations, recursion, and proofs.

3 Success geometry in arithmetic decoding

The contracts determine which couplings are allowed; success geometry explains their effects. Independent path strata cannot reduce batch success probability; lattices and fresh token strata can. We represent correct answers as interval regions to prove this guarantee and give counterexamples. Two-step comparisons then show gains and losses across decoders.

In this section, all lanes share a base decoder P and the same success rule. Fix a deterministic token ordering at each prefix. The arithmetic decoder $D : [0, 1) \to \mathcal{Y}$ maps interval points to complete outputs in $\mathcal{Y}$; for uniform U, $D(U) \sim P$. Let $A \subseteq [0, 1)$ be the set of points that decode to correct answers. Its total length, $\lambda(A) = s$, equals the individual success probability. Ordering the tokens differently rearranges this set while preserving its total length.

3.1 Independent path strata protect fixed batch subsets

Independent path strata protect fixed lane subsets and expected rewards with diminishing returns from successes. Both comparisons are relative to IID, averaged over sampler randomization with a shared decoder and success rule.

For the full batch, independence makes the probability of total failure a product of stratum-specific failure probabilities. With their average fixed, this product is largest when every stratum has the same failure probability, giving the IID value.

To make the argument precise, partition $[0, 1)$ into K equal strata, sample independently within each stratum, and uniformly permute the points over named lanes. Let $q_j = \mathbb{P}(U \in A \mid U \text{ lies in stratum } j)$, so $K^{-1} \sum_j q_j = s$.

A fixed subset of named lanes receives a uniformly chosen subset of strata. Averaging failure products over these possible assignments extends the comparison to smaller budgets.

Batch success exemplifies diminishing returns: only the first correct answer matters. More generally, let $\phi(n)$ reward n successes. Diminishing returns means that each additional success contributes no more than the previous one, expressed by discrete concavity: $\phi(n + 2) - 2\phi(n + 1) + \phi(n) \leq 0$. The theorem states both comparisons.

Theorem 1 (Path-stratification protection)**.** *Independent path stratification preserves every named path marginal.*

(i) Fixed-subset success. *For any fixed set of* $1 \leq m \leq K$ *named lanes,*

$$\mathbb{P}(\textit{pass@}m) = 1 - \frac{e_m(1 - q_1, \ldots, 1 - q_K)}{\binom{K}{m}} \geq 1 - (1 - s)^m, \tag{5}$$

where $e_m(b_1, \ldots, b_K) = \sum_{|J|=m} \prod_{j \in J} b_j$ *is the* m*th elementary symmetric polynomial.*

(ii) Success-count rewards. *If* N *is the full-batch success count and* $B \sim \mathrm{Binomial}(K, s)$ *is the success count under* K *IID draws, then every discrete-concave* $\phi : \{0, \ldots, K\} \to \mathbb{R}$ *satisfies* $\mathbb{E}\phi(N) \geq \mathbb{E}\phi(B)$*; the inequality reverses for discrete-convex* ϕ*.*

The guarantee concerns expected success for a lane subset fixed independently of outputs, not individual realized batches. Pass@1 is exactly preserved. The construction requires independent uniform offsets in fixed *path* strata, not a shared shift or strata redrawn after random prefixes.

Equal success probabilities across strata match IID; differences can produce a gain (Appendix Proposition 6). The proof uses Maclaurin's inequality and the Bernoulli-sum ordering of Hoeffding (1956); see Appendix A.5.

3.2 Lattices and fresh token strata can reduce pass@K

Lattices and fresh token strata can fail for different reasons. Lattice points can succeed or fail together, depending on overlap among shifted copies of A. Fresh token strata repartition after random prefixes, losing the fixed-path structure. We first calculate lattice coverage; all translations below are modulo one, wrapping around the unit interval.

Proposition 2 (Exact lattice criterion)**.** *For a random-shift K-point lattice,*

$$\mathbb{P}(\textit{pass@}K) = \lambda\left(\bigcup_{r=0}^{K-1}(A - r/K)\right). \tag{6}$$

For $K = 2$, its gain over IID is $s^2 - \lambda(A \cap (A - 1/2))$. The lattice helps, ties, or hurts depending on whether this half-period success overlap is below, equal to, or above s^2. In particular, $A = [0, 1/4) \cup [1/2, 3/4)$ gives pass@2 of $1/2$, below IID's $3/4$.

A fair two-step binary decoder realizes the counterexample when success means that the second token is zero and token zero is ordered first. Reversing the second-token order only after prefix one changes the success set to $[0, 1/4) \cup [3/4, 1)$. This set is disjoint from its shift by $1/2$, so lattice pass@2 becomes 1. Every path still has probability $1/4$, and IID pass@2 remains $3/4$. Ordering alone can therefore reverse the lattice's effect.

Fresh token stratification satisfies the strong contract in Eq. (2). However, repartitioning the next-token interval after random prefixes loses the fixed-stratum structure of Theorem 1.

Proposition 3 (Fresh token strata can hurt)**.** *For $K = 2$, let the first token be a fair binary token. At the second token, let success probability be $1/2 + \delta$ after prefix zero and $1/2 - \delta$ after prefix one, with $0 < \delta < 1/2$. Fresh independently jittered token strata with increasing-probability token order satisfy the strong contract and preserve pass@1 at $1/2$, but yield*

$$\mathbb{P}(\textit{pass@2}) = 1/2 + \delta. \tag{7}$$

This is below IID's $3/4$ when $\delta < 1/4$; at $\delta = 0.01$, it is 0.51.

The first round forces opposite prefixes. Increasing-probability order places successful token intervals at opposite ends of the two lanes' sampling intervals. Averaging over lane-to-stratum assignments gives simultaneous failure $(1 - 2\delta)/2$, despite exact conditional marginals. Appendix A.5 gives the calculations and proofs.

3.3 Exact two-step comparisons

Exact two-step comparisons show that average gains can coexist with losses on individual decoders. The strong contract can also reduce attainable success, while independent path strata remain protected.

For each of 20,000 decoders, we draw p, a_0, a_1 independently from $\mathrm{Beta}(1.5, 1.5)$. The first token is $Z \sim \mathrm{Bernoulli}(p)$; subsequent success satisfies $S \mid Z = z \sim \mathrm{Bernoulli}(a_z)$. We compute weak and strong pass@2 optima and integrate over code intervals to obtain the practical samplers' success probabilities. Each sampled decoder's pass probabilities are computed exactly (Appendix A.6).

IID averages 0.711 pass@2, while weak and strong optima average 0.836 and 0.826. The strong optimum is lower in 25.5% of instances (Figure 2; further gap statistics in Appendix A.6).

For the practical samplers, average gains can conceal losses on individual instances. Independent path strata never underperform IID. Lattices average 0.799 but lose on 15.2% of instances; fresh token strata average 0.765 and lose on 16.8%. The gap to the optimal values shows room to improve these sampling constructions.

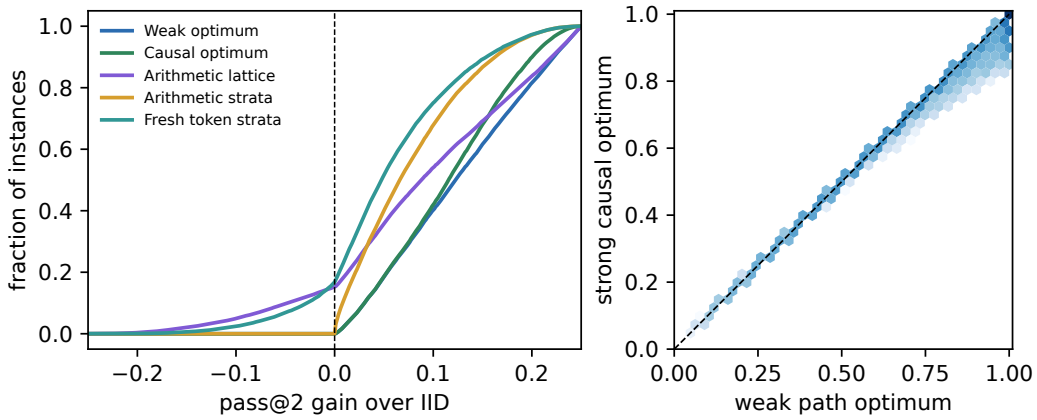

Figure 2: Exact pass@2 comparisons for 20,000 two-step decoders. **Left:** empirical cumulative distributions of gain over IID; negative values show underperformance on individual decoders, not estimation error. Independent path strata never underperform IID. **Right:** weak and strong optimal pass@2 values; the strong optimum is strictly lower in 25.5% of instances. The diagonal marks equality.

4 Language-model evidence across two regimes

The selected couplings show no supported gain on open-ended math, and the path-strata follow-up is inconclusive. The lattice improves success on Countdown. A separate Countdown study finds a moderate association between estimated geometry and measured gains. The four studies use separate evaluation populations; their differences do not explain the contrast between regimes. The experiments evaluate finite-precision implementations of the ideal constructions.

4.1 Do the selected couplings help on open-ended math?

Neither selected coupling yields a supported pass@8 improvement. Marginal accuracy meets equivalence, but passing this test and reducing text duplication do not establish better batch success. We compare a random-shift lattice and fresh token strata with IID; Arithmetic-IID controls for the arithmetic decoder.

Selection maximizes pooled pass@8 on 32 level-4/5 MATH training problems per model. Using the last number when a capped response lacks an explicit answer selects the lattice. Confirmation instead requires explicit answers from capped responses. This rule favors independent path strata, evaluated separately (Appendix B.1).

Design and inference. We evaluate Qwen2.5-3B-Instruct and Qwen2.5-Math-7B-Instruct (Qwen, 2025; Yang et al., 2024), abbreviated Qwen-3B and Qwen-Math-7B. We test both models on all 500 MATH-500 problems (Hendrycks et al., 2021; Lightman et al., 2023) and a fixed 500-problem GSM8K test subset (Cobbe et al., 2021). All evaluation problems are held out from selection.

Methods share prompts, stopping and grading, temperature 0.7, top-p 0.8, top-k 20, $K = 8$, and a 768-token cap. Capped responses without explicit answers are counted as incorrect; extraction and grading details are in Appendix B.1.

Each prompt supplies one batch per method. Direct pass@8 differences use prompt-paired bootstrap intervals and sign-flip tests, with Holm correction over the two coupled methods within each model–dataset block. Pass@1 equivalence uses a predeclared ± 0.03 margin; it tests marginal accuracy, not equality of complete-answer distributions. Appendix B.2 details inference and sensitivity analyses.

Results. All eight pass@8 intervals include zero (Table 2). An exact paired-binary sensitivity analysis also finds no supported gain (Appendix B.7). The estimates do not rule out positive gains: the Qwen-3B/MATH-500 token-strata interval extends to 0.050. Qwen-Math-7B/GSM8K has little

Table 2: Open-ended math, 500 prompts per block; the IID column reports pass@8. Δ_8 is the pass@8 difference from IID, with paired 95% bootstrap CIs. All eight intervals include zero; within-block Holm-adjusted $p \geq 0.659$. Full results and Arithmetic-IID controls appear in Table 7.

Model	Dataset	IID	Lattice Δ_8 [95% CI]	Token strata Δ_8 [95% CI]
Qwen-3B	MATH-500	0.670	0.014 $[-0.018, 0.046]$	0.018 $[-0.014, 0.050]$
	GSM8K	0.930	0.004 $[-0.014, 0.022]$	-0.010 $[-0.030, 0.010]$
Qwen-Math-7B	MATH-500	0.808	0.010 $[-0.008, 0.028]$	-0.006 $[-0.022, 0.010]$
	GSM8K	0.982	0.000 $[-0.006, 0.006]$	0.002 $[0.000, 0.006]$

room to improve over IID. MATH-500 truncation rates are 0.211–0.247, so these findings concern the capped decoder.

Arithmetic-IID and both couplings meet pass@1 equivalence in all four model–dataset settings: twelve comparisons. The lattice reduces Qwen-Math-7B text duplication without materially changing answer uniqueness (Appendix B.6).

4.2 Does path-stratification protection yield a detectable gain?

Independent path strata yield no detectable gain: one model ties IID and the other has a small negative difference. The follow-up uses 819 disjoint GSM8K problems with the same decoding and grading settings, specified before inspecting primary outcomes.

Pass@8 differences are -0.005 for Qwen-3B (95% CI $[-0.022, 0.011]$) and 0.000 for Qwen-Math-7B ($[-0.006, 0.006]$); both models meet the pass@1 equivalence criterion (Appendix Table 8). Theorem 1 guarantees nonnegative expected gain under ideal execution, not a strictly positive gain or a nonnegative difference in each finite evaluation.

4.3 Can a compact task exhibit a positive effect?

The lattice improves success on compact Countdown while meeting the marginal-accuracy equivalence criterion. Each item presents four arithmetic expressions using the same inputs. Exactly one reaches the target, as checked using exact rational arithmetic. A disjoint development study selects the $K = 8$ lattice (Appendix B.4).

Before evaluation, we fix Qwen-3B, temperature 1.0, top-p 0.95, and answer-only outputs capped at 24 tokens. We evaluate 512 reserved prompts with four batches per prompt and method. We average batches within prompt for paired inference; lattice–IID is the sole primary comparison, with Arithmetic-IID as a control. Pass@1 uses the same equivalence margin.

Pass@8 is 0.522 for IID and 0.573 for the lattice, a gain of 0.051 (95% CI $[0.038, 0.066]$, $p \approx 10^{-5}$; Table 3). The lattice meets pass@1 equivalence, with an estimated difference of 0.001. Arithmetic-IID shows no supported pass@8 difference from IID. Differences in task, output channel, and sampling settings prevent identifying the cause of the contrast with open-ended math.

4.4 Does estimated geometry predict independent effects?

Estimated geometry is moderately associated with independently measured Countdown gains. This establishes neither accurate gain predictions nor a rule for deciding when to use the lattice. We label code points by their decoded answers' correctness and compare predictions with gains from independent batches.

The geometry study uses a third, disjoint set of 128 Countdown prompts. Each prediction uses four independently shifted 64-point grids. Taking every eighth point partitions a grid into eight interleaved eight-point lattices. The fraction containing a correct answer estimates lattice pass@8. We subtract an IID estimate based on the grid's individual success rate and average over the four grids. Evaluation uses 16 fresh $K = 8$ batches per prompt and method with independent seeds. Appendix B.5 gives the estimator.

Table 3: Countdown confirmation ($K = 8$). Δ_8 and Δ_1 are pass@8 and lane-mean pass@1 differences from IID. Parse is the fraction of answers yielding an extracted A–D choice label. The lattice meets the predeclared ± 0.03 pass@1 equivalence margin. Sampling and inference are described in Section 4.3.

Method	pass@8	Δ_8	95% CI	p	Δ_1	Parse
IID	0.522	–	–	–	–	0.887
Arithmetic-IID	0.523	0.001	$[-0.012, 0.015]$	0.891	-0.002	0.880
Path lattice	0.573	0.051	$[0.038, 0.066]$	1.0×10^{-5}	0.001	0.883

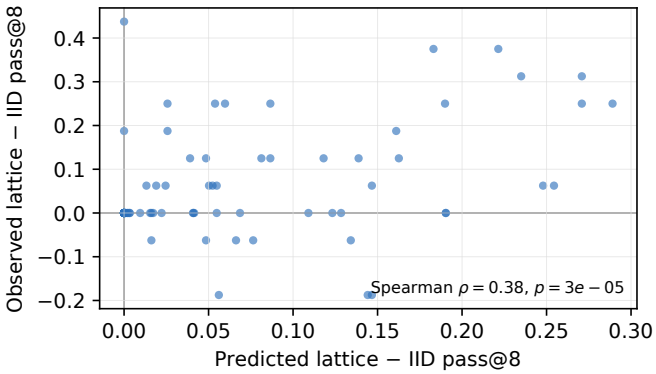

Figure 3: Geometry predictions and independent Countdown gains: 128 prompts, 256 graded samples per prediction, and 16 batches per prompt and method. Some points coincide. Association does not establish accurate gain predictions.

Spearman's rank correlation, $\rho = 0.376$, meets the predeclared positive-correlation criterion (one-sided permutation $p = 3.0 \times 10^{-5}$; Figure 3). Estimating IID success from the measured individual accuracy can bias predicted gains upward.

For predicted gains of magnitude at least 0.02, predictions of whether the lattice helps or hurts are no more accurate than always predicting a gain. The probe uses 256 graded samples per prompt, with no established advantage over simpler predictors or compute savings. Appendix B.5 reports calibration and sign diagnostics.

5 Limitations

Evidence from two checkpoints, two math benchmarks, and synthetic four-choice Countdown does not establish a universal sampler ranking. Pretraining exposure is unknown and grading remains imperfect. One batch per math prompt leaves variability across repeated batches unmeasured.

Exact guarantees concern ideal arithmetic, whereas implementations use finite precision. Geometry was measured only on Countdown. Task, output, and sampling differences prevent attributing the contrast with open-ended math to geometry. The diagnostic requires labeled samples and has not demonstrated an advantage over simpler predictors.

6 Conclusion

Coupled decoding separates individual fidelity from collective success. Preserving conditional next-token laws can reduce attainable success relative to preserving complete-path laws alone; neither contract guarantees improvement. Independent path strata protect fixed subsets under a shared decoder and success rule, while lattices and fresh token strata can help or hurt. The selected couplings show no supported gain on open-ended math. On Countdown, the lattice improves success, and estimated geometry is moderately associated with measured gains.

AI USE STATEMENT

The author developed the research idea and proofs. AI tools provided coding, debugging, writing, and editing support, including polishing the exposition and presentation of the proofs. The author reviewed and verified the resulting material and retains responsibility for the research questions, methods, proofs, experiments, interpretation, and final content.

ETHICS STATEMENT

This work studies coupled autoregressive decoding using public mathematical benchmarks and synthetic tasks. It does not involve human-subject experiments or the collection of private data. The guarantees concern marginal fidelity and batch success under specified assumptions; they should not be interpreted as guarantees of overall model safety. Changing dependence can also change the probability that a batch contains a harmful output, even when individual output distributions are preserved.

REPRODUCIBILITY STATEMENT

Appendix A gives complete proofs and finite-state calculations; Appendix B gives experimental conditions, inference, and supplementary results. The accompanying code provides sampling, generation, grading, analysis, and dataset-generation tools, along with model revisions and retained numerical summaries. It supports regeneration of the reported tables from those summaries and experiment reruns after obtaining the required assets. Most raw language-model generation records are not included, so the package does not support independent regrading or complete statistical reanalysis of those runs.

REFERENCES

Beatrice Acciaio, Daniel Kršek, and Gudmund Pammer. Multicausal transport: Barycenters and dynamic matching. *SIAM Journal on Financial Mathematics*, 16(3):1104–1138, 2025. doi: 10.1137/24M1634096.

Avinash Amballa, Aditya Parashar, Aditya Vikram Singh, Jinlin Lai, and Benjamin Rozonoyer. Quasi-random multi-sample inference for large language models. In *ICLR Workshop on Frontiers in Probabilistic Inference: Learning Meets Sampling*, 2025.

Julio Backhoff, Mathias Beiglböck, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017. doi: 10.1137/16M1080197.

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

Nina L. Corvelo Benz, Stratis Tsirtsis, Eleni Straitouri, Ivi Chatzi, Ander Artola Velasco, Suhas Thejaswi, and Manuel Gomez-Rodriguez. Evaluation of large language models via coupled token generation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2215–2223. PMLR, 2026.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Stephan Eckstein and Gudmund Pammer. Computational methods for adapted optimal transport. *The Annals of Applied Probability*, 34(1A):675–713, 2024. doi: 10.1214/23-AAP1975.

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

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956. doi: 10.1214/aoms/1177728178.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 2019.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. QuasiMoTTo: Quasi-Monte Carlo test-time scaling. *arXiv preprint arXiv:2607.01179*, 2026.

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. *arXiv preprint arXiv:2305.20050*, 2023.

Clara Meister, Martina Forster, and Ryan Cotterell. Determinantal beam search. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6551–6562. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.512.

Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12340.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic sampling: Parallel diverse decoding for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35120–35136. PMLR, 2023.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. URL `https://arxiv.org/abs/2409.12122`.

A Proofs and finite-state calculations

The proofs establish three distinctions used in the main text: conditional fidelity is stronger than path fidelity, that additional constraint can reduce attainable utility, and marginal preservation alone does not ensure better batch success. We first fix the information available to the sampler. We then prove the contract comparisons and entropy identities before turning to the geometric guarantees, counterexamples, and exact finite-state calculations.

A.1 Notation and conditional scope

Conditioning on completed joint rounds is sufficient to recover each lane's original path law. It does not guarantee fidelity after arbitrary partial revelation within a round. The notation below makes this scope explicit; the short conditioning argument then proves the implication.

Table 4 collects the notation used in the theoretical results. The general contract results allow different base laws P_r. Here $\Pi(\nu_1, \dots, \nu_K)$ denotes all joint laws with the stated marginals. We use natural-log relative entropy $\mathrm{KL}(\mu\|\nu) = \sum_y \mu(y) \log(\mu(y)/\nu(y))$ and total variation $\mathrm{TV}(\mu, \nu) = \frac{1}{2}\sum_y |\mu(y) - \nu(y)|$, with zero-mass entropy terms set to zero. The reference law is positive on the working support. The geometry guarantees use a common base law, arithmetic ordering, and success rule across lanes.

If Eq. (2) holds, conditioning further down to the lane's own prefix gives

$$\Pi(X_t^r \in \cdot \mid X_{<t}^r) = \mathbb{E}_\Pi[q_{r,t}(\cdot \mid X_{<t}^r) \mid X_{<t}^r] = q_{r,t}(\cdot \mid X_{<t}^r).$$

Table 4: Notation. A history h contains completed synchronous rounds only. In the geometry results, the token ordering is fixed before sampling.

Symbol	Definition
K, T, r, t	Number of named lanes, finite horizon, lane index, and joint-round index.
X_t^r, Y^r, P_r	Lane-r token at round t, complete stopped path, and its base law.
$q_{r,t}(\cdot \mid h^r)$	Filtered base next-token law, with absorbing stopped states.
$\mathcal{F}_{t-1}, h$	Information in $X_s^{1:K}$ for $s < t$, and a realized history.
$Q, \mathfrak{S}, \mathfrak{C}$	Product path law, weak coupling set, and strong coupling set.
$G, \varepsilon, L, \mathrm{OPT}_{\mathcal{A}}$	Bounded batch utility, positive entropy weight, regularized objective, and its optimum over $\mathcal{A}$.
D, A, s, λ	Arithmetic decoder, successful code set, its probability, and Lebesgue measure on the unit circle.
q_j, b_j, μ	Stratum-j success probability, failure probability $1 - q_j$, and mean failure probability $1 - s$.
$\mathcal{D}, v, b_{\max}$	Log-dispersion of failure masses, their variance, and their maximum.
N, ϕ, Δ_m	Full-batch success count, a function of that count, and the pass@m gain for a fixed subset of a K-stratum design.

Induction over t yields $Y^r \sim P_r$. This argument conditions on completed rounds. It makes no claim after a current-round token or the current jitter, shift, or permutation is exposed.

A.2 A WEAK COUPLING THAT IS NOT STRONG

The example preserves both complete-path marginals while changing a next-token conditional after the joint first round is observed. The key step is to restrict the uniform code point to the set compatible with that observation, then compute the next-token probability on the restricted set.

Let the two code points be U and $(U + 1/2) \bmod 1$, with U uniform. Both first tokens choose the interval of mass 0.7 exactly when $U \in [0, 0.2) \cup [0.5, 0.7)$. Conditional on this event, U is uniform on a set of total length 0.4. The first lane's residual is $U/0.7$. It lies below 0.8 on the entire first interval and on $[0.5, 0.56)$ in the second, giving conditional probability

$$\frac{0.2 + 0.06}{0.4} = 0.65.$$

Before conditioning on the joint first round, each code point is marginally uniform, so each lane has the base path distribution. The failure concerns precisely the second-round law given the completed first round, and witnesses $\mathfrak{C} \subsetneq \mathfrak{S}$.

A.3 STATIC PASS BOUNDS AND A STRICT CAUSAL GAP

Path marginals allow successes to be nested or spread around a common probability space, attaining both extreme union probabilities. The strict gap then arises because the strong contract restricts the couplings available at intermediate prefixes. We prove the static bounds first and solve the two-step conditional problem by optimizing its single free parameter.

Let lane r succeed with probability s_r.

Proposition 4 (Sharp weak pass range)**.** *Every weak coupling satisfies*

$$\max_r s_r \;\le\; \mathbb{P}(\textit{at least one success}) \;\le\; \min\Big\{1, \sum_r s_r\Big\}, \tag{8}$$

and both endpoints are attainable.

Proof of Proposition 4. The lower bound follows because the union contains each success event; the upper bound is the union bound, capped at one. To attain the lower endpoint, use nested events $[0, s_r)$ on a common uniform unit circle. To attain the upper endpoint, place arcs of lengths s_r consecutively around the circle. Their union has measure $\min\{1, \sum_r s_r\}$. For either construction, sample lane r from P_r conditional on its assigned success indicator. This lifts the indicator coupling to complete paths without changing any marginal. Conditional branches of zero probability are omitted. The lifted coupling need not satisfy the strong contract. □

Table 5: The contracts preserve different objects. The strong guarantee is conditional on completed synchronous rounds, not on partial revelation of the current round. Both contracts preserve every named complete-path law.

	Weak path marginal	Strong synchronous conditional
Additional requirement	None beyond $Y^r \sim P_r$	Eq. (2) at each round boundary
Optimization problem	Static path transport	Adapted, multicausal transport
Representative	Random-shift path lattice	Fresh token strata
Conditional audit	Not implied	At completed-round histories
Static pass ceiling	Attainable	May be unattainable

Proof of Proposition 1. Let each lane's first token equal one with probability $3/4$. Set terminal success probabilities to $a_0 = 1/20$ and $a_1 = 13/20$. Each lane's success rate is

$$s = (1/4)(1/20) + (3/4)(13/20) = 1/2.$$

Proposition 4 gives weak optimal pass@2 equal to one. For a strong first-round coupling, write $\theta = \mathbb{P}(X_1^1 = X_1^2 = 1) \in [1/2, 3/4]$. The masses of the four prefix pairs $(0,0), (0,1), (1,0), (1,1)$ are

$$\theta - 1/2, \quad 3/4 - \theta, \quad 3/4 - \theta, \quad \theta.$$

Conditional on first tokens (x, y), the optimal terminal success union is $\min\{1, a_x + a_y\}$, giving $1/10, 7/10, 7/10, 1$ for these four pairs. Thus the optimal continuation value at a given θ is

$$(\theta - 1/2)/10 + 2(3/4 - \theta)(7/10) + \theta = 1 - 3\theta/10.$$

The maximum is $17/20$, attained at $\theta = 1/2$. The conditional terminal couplings attain each stated union value, so this is an attainable strong optimum, not only an upper bound. □

A.4 Entropy-regularized optima

Both entropy-regularized problems have unique positive optimizers. Their objective regret—the gap between the optimal objective and that of another feasible coupling—is exactly a scaled KL divergence. These identities establish Corollary 1: the cost of imposing the strong contract equals a scaled KL divergence between the two optimizers. The strong problem also admits a backward recursion over completed histories.

We express the constraints as affine equalities, use entropy to establish positivity and uniqueness, and derive the objective-gap identities from stationarity. The chain rule then yields the recursion.

Throughout this subsection, the working space is the finite product of the supports of P_r, Q is positive on that space, G is bounded, and $\varepsilon > 0$. Positivity of an optimizer refers to this working support. The marginal potentials u_r below are functions of lane r's path that enforce its prescribed distribution. In the strong problem, $V_t(h)$ is the best expected terminal batch score minus the remaining entropy penalties, given the completed history h before round t. In the recursion, X denotes the next joint token vector, and (h, X) is the history after that round.

Proposition 5 (Weak entropic optimum)**.** *The weak problem has a unique positive optimizer of the form*

$$\Pi_w^*(y) = Q(y) \exp\left\{ \frac{G(y) - \sum_r u_r(y^r)}{\varepsilon} \right\}, \tag{9}$$

for marginal potentials u_r. Every weak-feasible Π satisfies

$$L(\Pi_w^*) - L(\Pi) = \varepsilon \mathrm{KL}(\Pi \| \Pi_w^*). \tag{10}$$

Theorem 2 (Strong entropic optimum)**.** *The strong class is a nonempty compact convex polytope. Its optimizer Π_c^* is unique and positive, and every strong-feasible Π obeys*

$$L(\Pi_c^*) - L(\Pi) = \varepsilon \mathrm{KL}(\Pi \| \Pi_c^*). \tag{11}$$

Relative entropy decomposes over joint-round kernels. With terminal value $V_{T+1} = G$, the optimal continuation value satisfies

$$V_t(h) = \max_{\pi \in \Pi(q_{1,t}(\cdot|h^1), \ldots, q_{K,t}(\cdot|h^K))} \left\{ \mathbb{E}_\pi V_{t+1}(h, X) - \varepsilon \mathrm{KL}\left(\pi \middle\| \bigotimes_r q_{r,t}(\cdot \mid h^r) \right) \right\}. \tag{12}$$

In particular, $V_1(\varnothing) = \mathrm{OPT}_{\mathfrak{C}}$.

Proof of Proposition 5. Existence and positivity. On the finite product support, the coupling set is compact and contains the strictly positive law Q. The objective is strictly concave, so the optimizer is unique. It is positive: mixing any boundary candidate with Q gives an infinite favorable entropy derivative at a zero coordinate, while the utility derivative is finite.

Optimizer form. Assign a multiplier $u_r(y^r)$ to each marginal constraint. The stationarity equation is

$$G(y) - \varepsilon\{1 + \log(\Pi(y)/Q(y))\} - \sum_r u_r(y^r) = 0.$$

Absorb the common constant into one potential to obtain Eq. (9).

Objective-gap identity. For every weak-feasible Π, fixed marginals make $\mathbb{E}_\Pi \sum_r u_r(Y^r)$ independent of Π. Substituting Eq. (9) into $\varepsilon\mathrm{KL}(\Pi\|\Pi_w^*)$ gives

$$\begin{aligned}\varepsilon\mathrm{KL}(\Pi\|\Pi_w^*) &= \varepsilon\mathrm{KL}(\Pi\|Q) - \mathbb{E}_\Pi G + \mathbb{E}_\Pi \sum_r u_r(Y^r) \\ &= L(\Pi_w^*) - L(\Pi),\end{aligned}$$

where the final constant follows by setting $\Pi = \Pi_w^*$. □

Proof of Theorem 2. Linear constraints. Represent a law on the finite path space by a vector z. Let $z(h, x)$ denote its induced next-prefix mass, with every later suffix marginalized out. Write x^{-r} for the next tokens of all lanes other than r. For each supported history h, lane r, and token i, the strong contract is equivalent to

$$\sum_{x^{-r}} z(h, i, x^{-r}) = q_{r,t}(i \mid h^r) \sum_x z(h, x). \tag{13}$$

At a zero-mass history both sides vanish. Prefixes excluded by the base support may be omitted. These are linear equalities; together with nonnegativity and unit mass they define a compact convex polytope. The positive product law Q is feasible. The induction in Appendix A.1 shows that these constraints also fix the path marginals.

Uniqueness and objective-gap identity. Strict concavity and the same mixture argument as in the weak problem give a unique positive optimizer. At this optimum, the objective gradient is a linear combination of the constraint rows. Every feasible distribution satisfies the same constraints, so its inner product with this gradient is constant. Since all distributions have unit mass, the expectation of $G - \varepsilon \log(\Pi_c^*/Q)$ is also constant across the strong feasible set. The calculation in the preceding proof then gives $L(\Pi_c^*) - L(\Pi) = \varepsilon\mathrm{KL}(\Pi\|\Pi_c^*)$.

Backward recursion. Write $\pi_t(\cdot \mid \mathcal{F}_{t-1})$ for the joint-round kernels induced by Π. The chain rule for relative entropy yields

$$\mathrm{KL}(\Pi\|Q) = \sum_{t=1}^{T} \mathbb{E}_\Pi \mathrm{KL}\left(\pi_t(\cdot \mid \mathcal{F}_{t-1}) \middle\| \bigotimes_{r=1}^{K} q_{r,t}(\cdot \mid X_{<t}^r)\right), \tag{14}$$

since the reference conditional is the displayed product. At a completed history h, the allowed next-round action is exactly $\Pi(q_{1,t}(\cdot \mid h^1), \ldots, q_{K,t}(\cdot \mid h^K))$. Backward induction with terminal payoff G and the additive entropy cost in Eq. (14) proves Eq. (12). □

Proof of Corollary 1. Because Π_c^* is weak-feasible, Eq. (10) applied to it gives

$$\mathrm{OPT}_{\mathfrak{S}} - \mathrm{OPT}_{\mathfrak{C}} = \varepsilon\mathrm{KL}(\Pi_c^*\|\Pi_w^*).$$

The same regret identity shows that maximizing L over $\mathfrak{C}$ is equivalent to minimizing the directed KL divergence to Π_w^*. □

A.5 Path-stratification guarantees and sampler failures

Independent path strata protect fixed-subset success and concave rewards of the total success count. Their proof rests on products of independent stratum failure probabilities and averaging over lane assignments. We then quantify the gain, derive the lattice overlap formula, and show why refreshing token strata can instead produce a loss. Finally, a total-variation argument explains why no dependent coupling improves every bounded batch utility.

Proof of Theorem 1. Uniformly permuting the strata makes each named code point uniform on $[0, 1)$, and therefore preserves the base path law.

Fixed-subset success. Write $b_j = 1 - q_j$. A fixed set of m named lanes receives a uniformly random m-subset of distinct strata. Conditional on that subset, the within-stratum draws are independent. Its all-failure probability is therefore

$$\binom{K}{m}^{-1} \sum_{|J|=m} \prod_{j \in J} b_j = \binom{K}{m}^{-1} e_m(b_1, \ldots, b_K).$$

Maclaurin's inequality bounds this expression by $(K^{-1} \sum_j b_j)^m = (1 - s)^m$. Taking complements proves Eq. (5), with equality at $m = 1$.

Success-count rewards. We now compare rewards based on the total number of successes. The key step is to equalize two success probabilities while keeping their sum fixed. The full success count is a sum of independent Bernoulli variables with probabilities $q_1, \ldots, q_K$; permuting them does not change the count. Hold the sum $p + r$ of two Bernoulli probabilities fixed and condition on the other variables' sum Z. Their contribution to the conditional expectation of ϕ is

$$\begin{aligned}(1 - p - r)\phi(Z) + (p + r)\phi(Z + 1) \\ + pr\{\phi(Z) - 2\phi(Z + 1) + \phi(Z + 2)\}.\end{aligned}$$

For discrete-concave ϕ, equalizing p and r increases pr and cannot increase this expectation. Repeated pairwise averaging converges to $(s, \ldots, s)$, giving $\mathbb{E}\phi(N) \geq \mathbb{E}\phi(\mathrm{Binomial}(K, s))$. For discrete-convex ϕ, the inequality reverses. This is the Bernoulli-sum comparison of Hoeffding (1956), applied to independent arithmetic path strata. □

Quantifying the stratification gain. The guarantee allows equality, so it does not promise a substantial gain on every problem. The gain depends on how unevenly success is distributed across strata. We quantify that dependence using their failure masses. Write $b_j = 1 - q_j$, $\mu = 1 - s$, and $\mathcal{D} = \log \mu - K^{-1} \sum_j \log b_j$. Set $\mathcal{D} = \infty$ if some $b_j = 0 < \mu$, and $\mathcal{D} = 0$ if $\mu = 0$. Let Δ_m denote the pass@m gain over IID for a fixed m-lane subset of the K-stratum design.

Proposition 6 (Quantitative stratification gain)**.** *For $K \geq 2$,*

$$\Delta_K = \mu^K - \prod_j b_j = \mu^K (1 - e^{-K\mathcal{D}}). \tag{15}$$

If $b_{\max} = \max_j b_j > 0$ and $v = K^{-1} \sum_j (b_j - \mu)^2$, then

$$\Delta_K \geq \mu^K \left\{ 1 - \exp\left(-\frac{Kv}{2b_{\max}^2} \right) \right\}, \qquad \Delta_2 = \frac{v}{K - 1}. \tag{16}$$

Thus differences in success probability across strata determine the benefit; equal probabilities match IID. For a fixed pair of lanes in the K-stratum design, the gain equals the across-stratum variance divided by $K - 1$. Token order can change this variance without changing s.

Proof of Proposition 6. The independent stratum draws have all-failure probability $\prod_j b_j$, while IID has probability μ^K. If all $b_j > 0$, the definition of $\mathcal{D}$ gives $\prod_j b_j = \mu^K e^{-K\mathcal{D}}$, proving Eq. (15). Jensen's inequality gives $\mathcal{D} \geq 0$; zero masses follow by continuity, with the all-success case handled by its stated convention.

On $(0, b_{\max}]$, $\log x$ is $1/b_{\max}^2$-strongly concave. Its quadratic upper support at μ, averaged over the b_j, gives

$$\frac{1}{K} \sum_j \log b_j \leq \log \mu - \frac{1}{2b_{\max}^2 K} \sum_j (b_j - \mu)^2.$$

The linear term cancels, so $\mathcal{D} \geq v/(2b_{\max}^2)$. Substitution proves the first bound in Eq. (16); limits again cover zero stratum failure masses.

A pair of named lanes receives a uniformly random pair of distinct strata. Its all-failure probability is

$$\binom{K}{2}^{-1} \sum_{i<j} b_i b_j = \frac{(\sum_j b_j)^2 - \sum_j b_j^2}{K(K-1)} = \mu^2 - \frac{v}{K-1}.$$

Subtracting from IID's failure probability proves $\Delta_2 = v/(K-1)$. The variance of failure masses equals that of success masses. □

Lattice overlap and failure. The lattice uses a shared shift rather than independent within-stratum draws. Its success therefore depends on overlap among shifted success regions.

Proof of Proposition 2. For points $(U + r/K) \bmod 1$, at least one succeeds exactly when the uniform shift U belongs to $\bigcup_{r=0}^{K-1}(A - r/K)$. This proves Eq. (6). At $K = 2$, inclusion–exclusion gives lattice pass probability $2s - \lambda(A \cap (A - 1/2))$, whereas IID gives $2s - s^2$. Their difference is the stated criterion. For $A = [0, 1/4) \cup [1/2, 3/4)$, both the set's measure and its half-period overlap are $1/2$. Either both lattice points succeed or neither does, giving pass@2 of $1/2$, compared with IID's $3/4$. □

Failure after refreshing token strata. Fresh token strata fail for a different reason: the success intervals depend on the prefixes reached in the previous round.

Proof of Proposition 3. The two first-round strata force opposite fair bits. Given prefix zero, success probability is $1/2 + \delta$, so increasing-probability order places failure in the lower interval $[0, 1/2 - \delta)$. Given prefix one, success is rarer and failure occupies $[1/2 - \delta, 1)$. If the lane with prefix zero receives the upper stratum in the second round, its failure is impossible. Under the other assignment, that lane fails with probability $1 - 2\delta$ and the lane with prefix one fails with probability one. Averaging over the two assignments gives simultaneous failure $(1 - 2\delta)/2$ and pass@2 $1/2 + \delta$. At each completed history, every named lane receives each stratum with equal probability and hence a uniform inverse-CDF input. The strong contract holds, and the path success marginal remains $1/2$. □

Limits beyond success-count utilities. No nonindependent coupling can improve every bounded batch utility. For any exact-marginal Π and the independent law Q,

$$\sup_{0 \le H \le 1} (\mathbb{E}_\Pi - \mathbb{E}_Q) H = \mathrm{TV}(\Pi, Q), \qquad \inf_{0 \le H \le 1} (\mathbb{E}_\Pi - \mathbb{E}_Q) H = -\mathrm{TV}(\Pi, Q). \tag{17}$$

This standard total-variation identity implies that every $\Pi \neq Q$ helps some utility and hurts another. The protected utilities in Theorem 1 are a specific class, not a universal ordering.

Proof of Eq. (17). An event attaining the positive part of the signed measure $\Pi - Q$ has probability difference $\mathrm{TV}(\Pi, Q)$. Its indicator attains the supremum; the complementary indicator attains the negative infimum. The variational characterization of total variation bounds every $[0, 1]$-valued utility by these two extremes. Because Π and Q share each lane marginal, unary utilities cancel. Any nonzero difference is therefore a dependence effect. □

A.6 Finite-state benchmark calculations

Every plotted pass probability is evaluated exactly for its sampled decoder. The strong optimum reduces to an affine objective in one first-round coupling parameter, so checking its endpoints suffices. The practical samplers are evaluated by interval integration. The random instance distribution therefore affects which decoders are compared, not the precision of their computed pass probabilities.

For each of the 20,000 instances, draw p, a_0, a_1 independently from $\mathrm{Beta}(1.5, 1.5)$. The complete-path success probability is $s = (1 - p)a_0 + pa_1$, and the weak optimal pass@2 is $\min\{1, 2s\}$. A strong first-round coupling is determined by

$$\theta = \mathbb{P}(Z^1 = Z^2 = 1) \in [\max(0, 2p - 1), p].$$

For a prefix pair (x, y), the optimal terminal union is $\min\{1, a_x + a_y\}$. The resulting objective is affine in θ, so checking the two endpoints gives the exact strong optimum. Instance selection uses seed 20260811. No entropic optimization is used for Figure 2; the plotted oracle values are unregularized pass@2 optima.

Path samplers. Partition $[0, 1)$ into first-token intervals of lengths $1-p$ and p, ordered by decreasing probability. Subdivide the interval for prefix z into failure and success intervals with relative lengths $1 - a_z$ and a_z, again in decreasing-probability order. The interval for path (z, i) has length $\mathbb{P}(Z = z)\mathbb{P}(S = i \mid Z = z)$. Let F be the union of the two failure intervals, and set $H_0 = [0, 1/2)$ and $H_1 = [1/2, 1)$. Then

$$p_{\mathrm{lat},2} = 1 - \lambda\big(F \cap (F - 1/2)\big),$$
$$p_{\mathrm{str},2} = 1 - \big[2\lambda(F \cap H_0)\big]\big[2\lambda(F \cap H_1)\big].$$

The first expression measures simultaneous failure of the shifted pair, with shifts modulo one. The second multiplies the independent failure probabilities within the two halves. Both are computed by summing interval intersection lengths.

Fresh token strata. Here the intervals are rebuilt after each prefix, using increasing-probability order. For a binary token with probability v of value one, let $I_i(v)$ be the interval assigned to value $i \in \{0, 1\}$. Define its probabilities under draws from the lower and upper halves by

$$\ell_i(v) = 2\lambda\big(I_i(v) \cap H_0\big), \qquad u_i(v) = 2\lambda\big(I_i(v) \cap H_1\big).$$

Averaging the two equally likely assignments of halves to lanes gives first-round prefix-pair probabilities

$$w_{xy} = \tfrac{1}{2}\{\ell_x(p)u_y(p) + u_x(p)\ell_y(p)\}, \qquad x, y \in \{0, 1\}.$$

Given prefixes (x, y), value zero in the terminal round means failure. Fresh independent draws and a fresh lane assignment therefore give

$$p_{\mathrm{tok},2} = 1 - \sum_{x,y\in\{0,1\}} w_{xy}\tfrac{1}{2}\{\ell_0(a_x)u_0(a_y) + u_0(a_x)\ell_0(a_y)\}.$$

Thus all three samplers are evaluated from interval lengths, without Monte Carlo estimation. In both ordering conventions, probability ties place value zero before value one.

IID averages 0.711 pass@2; the weak and strong optima average 0.836 and 0.826. Their mean gap is 0.0091, with maximum 0.158, and the strong optimum is strictly lower in 25.5% of instances (Figure 2).

B Experimental protocol and supplementary results

These details support the interpretation of the empirical findings. Method selection depends on the grading rule, and finite-precision execution does not itself certify the ideal marginal guarantees. We document those conditions first, then explain paired inference, present the full open-ended results, and describe compact-task development and the geometry estimator. The supplementary diagnostics show that reduced text duplication does not establish improved correctness, that the exact-test sensitivity leaves the primary conclusion unchanged, and that grader disagreement remains a source of uncertainty.

B.1 Method selection and numerical conditions

Method selection depends on answer extraction: the development rule selects the lattice, whereas the confirmation rule favors independent path strata. The reported numerical samplers also require qualification: the ideal marginal guarantees are not a certificate of full finite-precision path-law equality. The selection procedure and execution conditions below specify what was evaluated and which conclusions follow from it.

Method selection uses 32 level-4/5 MATH training problems per model at $K = 8$, giving 64 prompt–model batches per core method. The weak candidates are a random-shift lattice, independent

path strata, and scrambled Sobol. The strong candidates are fresh systematic and independently jittered token strata. Selection maximizes pooled direct pass@8, using duplicate rate and then throughput as tie-breakers. The selected methods are the lattice and fresh jittered token strata; IID and Arithmetic-IID are controls.

Scrambled Sobol uses randomized points designed to cover a unit cube evenly. Each point is assigned to a lane, and successive coordinates supply its successive tokens' inverse-CDF inputs. Assignment to lanes is uniformly randomized once per batch. Fresh systematic token strata instead use equally spaced points with one shared random offset per round, uniformly assigned to active lanes. As in fresh jittered strata, tokens are ordered by increasing probability; the difference is the shared offset rather than independent offsets within strata.

The development extraction rule allows a last-number fallback at the token cap. The confirmation rule does not, and is fixed before held-out generation. Applying the confirmation rule to the development data gives independent path strata pooled pass@8 of 0.500, compared with the lattice's 0.484; the strong-method selection is unchanged. The primary comparison uses the lattice selected under the development rule. The disjoint 819-prompt path-strata study is specified before primary outcomes are inspected and is analyzed separately.

Shared evaluation conditions. A lane stops on either model EOS token or a complete explicit answer marker, checked every eight generated tokens. A cap-truncated response without an explicit answer counts as a failure. GSM8K uses terminal numeric equivalence; MATH-500 uses Math-Verify on the extracted answer with a deterministic numeric/symbolic fallback. Each arithmetic residual remains attached to its lane, including when that lane is the only active one. Together with shared stopping rules, this is necessary to interpret the samplers as couplings of the same decoder.

Models run in bfloat16 on two RTX 4090 GPUs, with one resident model per device and serial jobs per device.

Numerical scope. Code points and residuals are stored in float64. Filtered probabilities are computed in float32 from bfloat16 model execution. The implementation uses finite pseudorandom draws and no random-bit refill. Exact marginal preservation is therefore a property of the ideal continuous-uniform, exact-arithmetic construction; full path-law equality for the numerical implementation is not established. Batch-dependent changes in model kernels also fall outside the assumption that each base kernel depends only on its own lane's prefix. Arithmetic-IID and pass@1 equivalence check selected observable consequences of the implementation, rather than certify its full distribution.

B.2 Endpoints, uncertainty, and multiplicity

Inference is paired at the prompt level. Repeated batches in the compact studies are averaged within prompt, rather than counted as independent problems. Success tests and marginal-accuracy equivalence tests answer different questions; neither an equivalence result nor nonsignificance certifies complete-distribution fidelity. We define the endpoints before giving the resampling and testing procedures.

Estimation and uncertainty. For a prompt and a method, direct pass@8 is the indicator that at least one of its eight generated lanes is correct. Lane-mean pass@1 is the mean of the eight correctness indicators. The primary and protected studies observe one batch per prompt and method. Method differences use matched prompts, not identical sampler seeds. In the Countdown studies, repeated batches are first averaged within prompt; prompt-level pairing is then used for inference.

Primary 95% confidence intervals use 10,000 prompt-clustered bootstrap replicates. The protected comparison and compact confirmation each use 10,000 paired prompt-bootstrap replicates.

For a paired binary difference in $[-1, 1]$, the distribution-free standard-error bound with 500 independent prompt units is $1/\sqrt{500} = 0.045$ (rounded). The reported bootstrap intervals exploit the observed pairing and are typically tighter. The one-batch studies use prompt-level replication; they do not separately estimate the randomization distribution at each fixed prompt. Nonsignificance is therefore not evidence that all smaller gains vanish.

Efficacy tests. Primary efficacy tests use paired sign flips with Holm correction across the lattice and token-strata comparisons within each model–dataset block. The protected comparison uses 100,000 sign-flip replicates. The compact confirmation uses a two-sided, 100,000-replicate sign-flip test for its sole primary lattice comparison.

Geometry association. The geometry study specifies a one-sided, 100,000-permutation test of positive Spearman correlation, with a predeclared $p < 0.05$ criterion. Its prediction and evaluation samples are independently randomized.

Testing assumptions. No post-generation exclusions, adaptive stopping, or outcome-dependent method reselection enter the reported comparisons. Monte Carlo sign-flip and permutation p-values include the plus-one correction; values near 10^{-5} are at the resolution of 100,000 replicates. Exact sign-flip calibration requires sign exchangeability of paired differences under the null, which is stronger than a zero-mean condition alone.

Marginal-accuracy equivalence. Table 6 reports every primary and protected pass@1 comparison against the predeclared equivalence margin ± 0.03. All twelve nonbaseline primary and both protected comparisons pass paired two one-sided tests (TOST), using the paired mean and standard error with a Student-t reference distribution and $n - 1$ degrees of freedom for n prompt units. The compact confirmation also uses paired TOST with the same margin. These tests concern lane-mean accuracy; neither a TOST result nor a small range of named-lane accuracies establishes the complete-distribution contract. The exact-marginal claims follow from the sampling constructions under their stated implementation assumptions.

Table 6: Lane-mean pass@1 differences from IID. The intervals are paired 95% prompt-bootstrap CIs; *Equivalent* reports the predeclared paired two one-sided tests (TOST) with margin ± 0.03. A 95% interval need not lie within this margin for TOST to pass. Primary blocks contain 500 prompts; the path-strata follow-up contains 819. This accuracy check is not an empirical test of the complete path law or every conditional marginal.

Method or model	Δ_1	95% CI	Equivalent
Qwen-3B / MATH-500			
Arithmetic-IID	-0.010	$[-0.026, 0.006]$	yes
Path lattice	-0.009	$[-0.025, 0.006]$	yes
Token strata	-0.010	$[-0.025, 0.006]$	yes
Qwen-3B / GSM8K			
Arithmetic-IID	-0.003	$[-0.021, 0.016]$	yes
Path lattice	-0.012	$[-0.032, 0.007]$	yes
Token strata	-0.006	$[-0.025, 0.013]$	yes
Qwen-Math-7B / MATH-500			
Arithmetic-IID	0.001	$[-0.008, 0.010]$	yes
Path lattice	-0.004	$[-0.013, 0.006]$	yes
Token strata	-0.005	$[-0.015, 0.005]$	yes
Qwen-Math-7B / GSM8K			
Arithmetic-IID	0.005	$[-0.001, 0.011]$	yes
Path lattice	-0.002	$[-0.006, 0.003]$	yes
Token strata	0.001	$[-0.004, 0.005]$	yes
Independent path strata / GSM8K remainder ($n = 819$)			
Qwen-3B	-0.005	$[-0.019, 0.009]$	yes
Qwen-Math-7B	-0.001	$[-0.005, 0.003]$	yes

B.3 Full open-ended efficacy results

The full results support the main-text finding of no statistically supported gain on the tested open-ended math problems. All eight primary coupled-method confidence intervals include zero; the protected path-strata follow-up is also inconclusive. Tables 7 and 8 report the primary and follow-up results, including controls, individual test values, and baseline accuracies.

Table 7: Open-ended math: directly observed pass@8 and prompt-paired differences from IID, with 500 prompts per model–dataset block. Intervals are 95% prompt-bootstrap CIs. p_H is the paired sign-flip p-value after Holm correction over the lattice and token-strata contrasts within each block. Arithmetic-IID is an implementation control, outside this testing family. All values are proportions.

Method	pass@8	Δ_8	95% CI	p_H
Qwen-3B / MATH-500				
IID	0.670	–	–	–
Arithmetic-IID	0.680	0.010	$[-0.020, 0.040]$	–
Path lattice	0.684	0.014	$[-0.018, 0.046]$	0.659
Token strata	0.688	0.018	$[-0.014, 0.050]$	0.659
Qwen-3B / GSM8K				
IID	0.930	–	–	–
Arithmetic-IID	0.922	−0.008	$[-0.028, 0.010]$	–
Path lattice	0.934	0.004	$[-0.014, 0.022]$	0.886
Token strata	0.920	−0.010	$[-0.030, 0.010]$	0.886
Qwen-Math-7B / MATH-500				
IID	0.808	–	–	–
Arithmetic-IID	0.810	0.002	$[-0.012, 0.018]$	–
Path lattice	0.818	0.010	$[-0.008, 0.028]$	0.764
Token strata	0.802	−0.006	$[-0.022, 0.010]$	0.764
Qwen-Math-7B / GSM8K				
IID	0.982	–	–	–
Arithmetic-IID	0.982	0.000	$[-0.006, 0.006]$	–
Path lattice	0.982	0.000	$[-0.006, 0.006]$	1.000
Token strata	0.984	0.002	$[0.000, 0.006]$	1.000

Table 8: Independent path-stratification follow-up on the disjoint 819-problem GSM8K remainder, with one $K = 8$ batch per prompt and method. Differences are relative to IID; intervals are paired 95% prompt-bootstrap CIs and p is the paired sign-flip value. Both pass@1 equivalence tests pass (Table 6).

Model	IID	Path strata	Δ_8	95% CI	p
Qwen-3B	0.934	0.929	−0.005	$[-0.022, 0.011]$	0.658
Qwen-Math-7B	0.977	0.977	0.000	$[-0.006, 0.006]$	1.000

B.4 Compact-task development

The compact study separates design selection from reserved confirmation. Development selects the $K = 8$ random-shift lattice, which then shows a positive effect on the reserved confirmation set (Table 3). The development, confirmation, and geometry studies use disjoint prompt sets.

Development uses Qwen-3B on 128 deterministic multiple-choice Countdown prompts, with two independently randomized batches per prompt, method, and $K \in \{4, 8\}$. The candidates are the random-shift lattice, independent path strata, and scrambled token Sobol, each compared with IID. Generation uses temperature 1.0, top-p 0.95, no top-k filter, and answer-only outputs capped at 24 tokens. Confirmation uses 512 disjoint prompts, and the geometry study uses another 128. Each question has four expressions using the same input multiset and exactly one correct answer under exact rational grading.

The selection rule operates at $K = 8$. A candidate must yield an extracted A–D label in at least 80% of outputs and meet pass@1 equivalence with margin ± 0.03. Among eligible candidates, the largest mean paired pass@8 gain over IID is selected. Confirmation proceeds only if that gain is at least 0.02.

All three candidates meet the eligibility criteria. IID development pass@8 is 0.5234; the lattice, path strata, and token Sobol have respective gains of 0.0352, 0.0156, and 0.0039. The lattice therefore qualifies for confirmation. Its development gain has a 95% prompt-bootstrap interval of $[0.0000, 0.0703]$ and a two-sided sign-flip $p = 0.0932$. At $K = 4$, the respective gains are 0.0117,

0.0000, and 0.0078. These development comparisons select the confirmation method; they do not establish efficacy. The positive effect in the main text comes from the separate reserved confirmation study.

B.5 What the geometry diagnostic estimates

The diagnostic estimates lattice coverage from shifted grids and subtracts an IID baseline estimated from individual accuracy. The grid average is unbiased for ideal lattice coverage, but the nonlinear baseline can bias the estimated gain upward. We first construct the estimator, then explain this bias and distinguish association, sign agreement, and calibration.

Constructing the prediction. The geometry study labels four independently shifted 64-point lattices per prompt. Within one grid, sort the points around the unit circle and let $b_i \in \{0, 1\}$ denote correctness at position i, for $i = 0, \ldots, 63$. The estimated individual success probability is $\widehat{s} = 64^{-1} \sum_i b_i$.

Taking every eighth point forms eight sublattices, each containing eight points. A sublattice succeeds if any of its points decodes to a correct answer. Averaging these eight success indicators gives the grid's estimate of lattice pass@8:

$$\widehat{p}_{\mathrm{lat},8} = \frac{1}{8} \sum_{r=0}^{7} \mathbf{1}\left\{ \max_{0 \le j < 8} b_{r+8j} = 1 \right\}.$$

After sorting, a fixed sublattice need not have a uniform shift. Choosing its index uniformly from the eight possibilities recovers the random-shift eight-point lattice law under the ideal common decoder. The displayed average therefore estimates lattice pass@8 unbiasedly, although its eight terms are dependent. This is the lattice subsampling structure of Li et al. (2026).

For each grid, subtract the estimated IID success probability $1 - (1 - \widehat{s})^8$ from $\widehat{p}_{\mathrm{lat},8}$. The final prediction averages these four gain estimates. Observed gains use the separate evaluation with 16 batches per prompt and method described in Section 4.4.

Why the estimated gain can be biased upward. The upward bias can arise from the estimated IID baseline, even when the lattice-coverage estimate is unbiased. Since $1 - (1 - s)^8$ is concave in s, a marginally unbiased $\widehat{s}$ yields a downward-biased estimate of IID pass@8. Subtracting this baseline gives an upward-biased estimated gain when the lattice term is unbiased.

Association with measured gains. The correlation reported in Section 4.4 assesses the finite-sample diagnostic against independently measured gains. It does not establish exact calibration of the population criterion. The diagnostic's incremental value beyond marginal success and answer concentration remains untested.

Sign agreement. The secondary sign statistic compares the signs of predicted and observed differences among prompts with absolute prediction at least 0.02. A zero observed difference counts as a mismatch to a nonzero prediction. All 45 selected predictions are positive; 27 observed differences are positive, 11 are zero, and 7 are negative. Thus, sign agreement is $27/45 = 60\%$, identical to always predicting a positive gain on this subset.

Calibration. Calibration concerns agreement between predicted and observed gain sizes. Mean predicted gain is 0.0420 and mean observed gain is 0.0366 (95% CI $[0.0200, 0.0552]$). The mean residual, observed minus predicted, is -0.0054 (95% CI $[-0.0215, 0.0106]$). This interval assesses the finite-sample diagnostic; it includes zero but does not establish exact calibration.

B.6 Diversity and generation measures

We now return to the open-ended math evaluations to examine repetition, answer diversity, and generation behavior. The lattice reduces text duplication in the specialist model without a supported improvement in correct-answer coverage. The measures below separate repetition, extracted-answer diversity, and generation behavior so that changes in one are not mistaken for changes in another.

Table 9 gives all reported secondary primary-study measures: normalized unique answers, whitespace-normalized text duplicate rate, explicit-answer rate, truncation rate, mean generated length, and throughput. Duplicate rate is the fraction of the $\binom{8}{2}$ unordered lane pairs whose full generated texts match after stripping outer whitespace and collapsing each internal whitespace run to one space. Normalized unique answers instead counts distinct nonmissing extracted answer strings divided by eight. Throughput is descriptive for the bfloat16, two-RTX-4090 setup; each model is resident on a separate GPU and its jobs are processed serially.

All twelve nonbaseline pass@1 comparisons meet the equivalence margin (Table 6). The lattice reduces Qwen-Math-7B text duplicate rates from 0.010 to 0.004 on MATH-500 and from 0.029 to 0.011 on GSM8K, while normalized answer uniqueness is nearly unchanged (Table 9).

Table 9: Secondary open-ended generation measures. Uniq. is normalized unique answers; Dup. is the whitespace-normalized text duplicate rate; Explicit and Trunc. are explicit-answer and truncation rates; Length is mean generated tokens per lane. Throughput is generated tokens per second in the stated hardware configuration. Values retain the reported precision; rounded zeros need not be exact zeros.

Method	Uniq.	Dup.	Explicit	Trunc.	Length	Tok./s
Qwen-3B / MATH-500						
IID	0.481	0.000	0.778	0.222	477.7	306.1
Arithmetic-IID	0.501	0.000	0.789	0.211	477.2	294.2
Path lattice	0.486	0.000	0.778	0.222	483.0	294.7
Token strata	0.499	0.000	0.783	0.217	480.4	297.8
Qwen-3B / GSM8K						
IID	0.538	0.000	0.999	0.002	252.4	299.0
Arithmetic-IID	0.545	0.000	0.998	0.002	251.5	289.0
Path lattice	0.540	0.000	0.999	0.001	254.2	287.7
Token strata	0.537	0.000	0.999	0.001	253.7	292.0
Qwen-Math-7B / MATH-500						
IID	0.144	0.010	0.766	0.234	512.1	324.6
Arithmetic-IID	0.141	0.012	0.759	0.241	517.0	319.3
Path lattice	0.142	0.004	0.753	0.247	516.4	314.4
Token strata	0.144	0.007	0.759	0.241	515.9	316.6
Qwen-Math-7B / GSM8K						
IID	0.143	0.029	0.998	0.003	300.1	319.7
Arithmetic-IID	0.142	0.030	0.996	0.004	300.4	315.5
Path lattice	0.145	0.011	0.997	0.003	301.5	314.3
Token strata	0.143	0.022	0.997	0.003	301.0	315.6

B.7 Binary-outcome sensitivity and named-lane diagnostics

An exact binary-outcome sensitivity analysis gives the same qualitative conclusion as the primary inference: no supported gain in the tested contrasts. Named-lane accuracy ranges provide an additional descriptive check, but do not establish exchangeability or conditional fidelity. We describe these two diagnostics separately.

The prospectively specified exact sensitivity treats discordant paired prompt outcomes as a binomial test (equivalently, exact McNemar). Table 10 reports every discordant gain/loss count. Holm adjustment is global across the eight proposed primary contrasts and, separately, across the two protected-model contrasts. Every adjusted p-value equals 1.000. This broader testing family is distinct from the within-block primary family in Table 7.

Table 10: Exact binary-outcome sensitivity analysis. Gains/losses are discordant prompt counts relative to IID. The exact McNemar (binomial) p-values are Holm-adjusted across all eight proposed primary contrasts, and separately across the two protected path-strata contrasts. All adjusted values are 1.000.

Method	Gains/losses	Adjusted p
Primary Qwen-3B / GSM8K		
Lattice	$12/10$	1.000
Token strata	$11/16$	1.000
Primary Qwen-3B / MATH-500		
Lattice	$39/32$	1.000
Token strata	$38/29$	1.000
Primary Qwen-Math-7B / GSM8K		
Lattice	$1/1$	1.000
Token strata	$1/0$	1.000
Primary Qwen-Math-7B / MATH-500		
Lattice	$13/8$	1.000
Token strata	$7/10$	1.000
Protected Qwen-3B / GSM8K rem.		
Path strata	$21/25$	1.000
Protected Qwen-Math-7B / GSM8K rem.		
Path strata	$3/3$	1.000

Table 11 reports the maximum-minus-minimum range of the eight named-lane pass@1 rates for all primary and protected methods, including IID. These are descriptive diagnostics of named-lane accuracy, not formal evidence of exchangeability or conditional fidelity.

Table 11: Named-lane accuracy diagnostic: maximum minus minimum of the eight named-lane pass@1 rates in each block. *Path coupling* denotes the lattice in the primary study and independent path strata on the GSM8K remainder. These descriptive ranges do not establish exchangeability or replace the paired equivalence tests.

Model / dataset	n	IID	Arith.-IID	Path coupling	Token strata
Qwen-3B / GSM8K	500	0.076	0.062	0.054	0.044
Qwen-3B / MATH-500	500	0.040	0.048	0.032	0.040
Qwen-Math-7B / GSM8K	500	0.006	0.008	0.016	0.020
Qwen-Math-7B / MATH-500	500	0.030	0.022	0.020	0.030
Qwen-3B / GSM8K rem.	819	0.059	–	0.035	–
Qwen-Math-7B / GSM8K rem.	819	0.016	–	0.017	–

B.8 Grading disagreements

The graders disagree more often on MATH-500 than on GSM8K, with different patterns across models. These discrepancies describe grading uncertainty; they are not adjudicated error rates. The comparison holds the extracted answer fixed so that it assesses grading rather than answer extraction.

The primary grader uses the terminal extracted answer, with cap-truncated responses lacking an explicit answer counted as failures. GSM8K is graded by numeric equivalence; MATH-500 uses Math-Verify with the deterministic numeric/symbolic fallback (Math-Verify version 0.9.0; `https://github.com/huggingface/Math-Verify`). Table 12 compares the fixed primary grade with the deterministic numeric/symbolic fallback on the *same* extracted terminal answer, without changing the primary outcome. It includes every primary and protected method. The table distinguishes answers accepted only by the primary grader from those accepted only by the fallback, showing the direction as well as the frequency of disagreement.

Table 12: Grading disagreements on the same extracted terminal answers. P-only and F-only count lanes accepted only by the primary grader or only by the deterministic numeric/symbolic fallback. Any/Rate gives their sum and the fraction of all lanes. Rates are rounded to three decimals, so a displayed zero can accompany a nonzero count. This descriptive comparison does not change the primary grades.

Method	Lanes	P-only	F-only	Any/Rate
Qwen-3B / GSM8K				
Arithmetic-IID	4000	1	0	1/0.000
Path lattice	4000	1	0	1/0.000
Token strata	4000	2	0	2/0.001
IID	4000	0	0	0/0.000
Qwen-3B / MATH-500				
Arithmetic-IID	4000	64	114	178/0.044
Path lattice	4000	83	123	206/0.051
Token strata	4000	77	109	186/0.046
IID	4000	70	104	174/0.043
Qwen-Math-7B / GSM8K				
Arithmetic-IID	4000	0	0	0/0.000
Path lattice	4000	0	0	0/0.000
Token strata	4000	0	0	0/0.000
IID	4000	0	0	0/0.000
Qwen-Math-7B / MATH-500				
Arithmetic-IID	4000	179	9	188/0.047
Path lattice	4000	188	14	202/0.051
Token strata	4000	178	22	200/0.050
IID	4000	177	20	197/0.049
Qwen-3B / GSM8K rem.				
Path strata	6552	1	0	1/0.000
IID	6552	2	0	2/0.000
Qwen-Math-7B / GSM8K rem.				
Path strata	6552	0	0	0/0.000
IID	6552	0	0	0/0.000