SEPARATING MARGINAL ERROR FROM RECOVERABLE DEPENDENCE IN PARALLEL-HEAD SPECULATIVE DECODING FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Speculative decoding commits several tokens per forward pass by checking a block a cheap drafter proposed. When that drafter is parallel heads, one per future position, they propose independently and miss the target's joint law for two confounded reasons: marginal error, and dependence no product can represent. We separate them exactly at depth two for two deployed head families whose first proposal is the target's own next-token head. Under that condition the deficit is second-marginal error plus a recoverable term, the total variation lost when the target's conditional mismatch is marginalised over the first position. The target's own dependence bounds that term, giving a normalised recoverable fraction that a certified estimator measures at vocabulary scale. On a fixed target, marginal error strongly predicts that fraction without determining it: at fixed prefixes the certified bounds establish that proposals with exactly equal second-marginal error can differ in recoverable fraction, and the target's own head does not behave like the perturbations that concentrate mass. Across three seeded runs on a frozen trunk, the calibration differs between prefix distributions of different first-position entropy, and a positive entropy-residual coupling appears with training. Because the recoverable term caps what dependence can add at fixed marginals, a small value rules out a large gain from that redesign before it is built.

1 INTRODUCTION

How fast a large language model serves depends on how many forward passes it makes, and speculative decoding (Leviathan et al., 2023) reduces that count: a cheap drafter proposes a block of future tokens, the target model checks the block in one pass, and the longest accepted prefix is committed, distributed exactly as the target alone would have produced it. One widespread drafter is a set of light heads attached to the target itself, each reading the same hidden state and predicting one future position, so the block costs one forward pass and no second model; Medusa adds such heads to a trained target (Cai et al., 2024) and Meta's multi-token-prediction models train them alongside it (Gloeckle et al., 2024). Because the heads never see one another's tokens they propose each position independently, while the target's own positions are not. Extra prediction heads are standard equipment in open-weight models: GLM-4.5 adds a multi-token-prediction layer to drive speculative decoding (GLM-4-5 Team, 2025), Kimi K3 serves one as its drafter (Kimi Team, 2026), and DeepSeek-V3 (DeepSeek-AI, 2024) and DeepSeek-V4 (DeepSeek-AI, 2026) keep a causal chain between the predicted positions instead of leaving them independent.

The engineer who owns such a drafter and wants more of each block to survive decides where the next months of work go: into making each head more accurate at its own position, or into rebuilding the drafter so the heads condition on one another, by feeding each head what the earlier heads sampled (Ankner et al., 2024), drafting autoregressively in the target's feature space (Li et al., 2024), or drafting the block jointly (Karimi et al., 2026; Tang et al., 2026). Accuracy does nothing when the heads are each right at their own position and their product still puts mass on combinations the target would not emit (Rusanovsky et al., 2026), a failure that grows worse at later positions (Oda et al., 2026); dependence cannot remove the floor a wrong marginal sets. The acceptance rate a drafter reports does not say which it is paying for, so the decision is made before the evidence that would settle it, and whoever decides wrong can spend months and move nothing.

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.

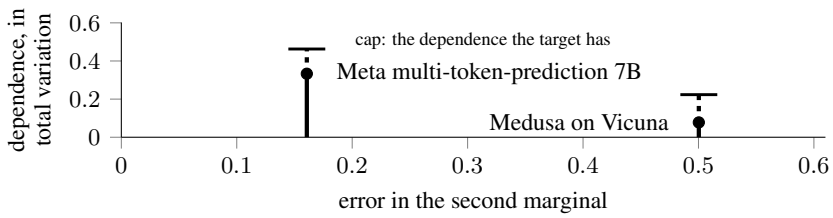


Figure 1: The two costs a parallel-head drafter pays at depth two, measured for two deployed families on the positions whose certificate for the cap is valid, 127 and 128 of 128. The horizontal coordinate is error in the drafter's second marginal, which a better second head reduces. The vertical is dependence in total variation: each bar is solid up to the part of the deficit no product of independent heads can represent, which the dot marks, and dashed on to the cap, the dependence the target itself has. The solid fraction is the share still recoverable, 0.3484 for Medusa and 0.7198 for the multi-token-prediction model.

This paper measures one side of that choice in advance. We work at depth two, the first two proposed positions, where the deficit of the head families studied here separates exactly into a marginal cost and a dependence cost. Both are measurable on the drafter that already exists, without training a second one. At fixed per-position marginals, no change to the joint arrangement can recover more than the recoverable term, so a small value rules out a large dependence-only gain before such a redesign is built. Medusa (Cai et al., 2024) and Meta's multi-token-prediction models (Gloeckle et al., 2024) both draw their first proposed position from the target's own next-token head, so the first marginal is correct by construction rather than by measurement, and Appendix A says what the records can and cannot check about that. Under that condition the distance from the target's joint law to the set of all laws with the drafter's two marginals collapses to a distance between second marginals, and the loss splits into a marginal term and a term dependence can recover. The general case needs a third term for the amount by which two separately plausible marginals cannot be arranged to match the target, and for these architectures it is zero.

The recoverable term is the total variation lost when the target's own conditional mismatch is averaged over the first position, a gap that exists because averaging inside a distance is not averaging outside it. Dividing it by the target's own dependence, which bounds it, leaves a fraction between zero and one that puts different targets on one scale: that half of a target's dependence is still recoverable at a head's marginals is comparable between systems, and an acceptance rate is not. Both terms involve the target's conditional distribution over 32000 tokens at every candidate first token, so each is reported as a certified interval.

Figure 1 puts both on one pair of axes: Medusa loses most of its deficit to its second marginal and Meta's multi-token-prediction model loses most of its to factorisation, so two shipped systems raise opposite hypotheses and neither can read the answer off the other.

Our contributions are:

1. When a parallel drafter takes its first position from the target's own next-token head, its depth-two deficit is exactly second-marginal error plus a recoverable term, with no third term, so the confound the deficit hides is separable. That statement, the coupling bound it is a special case of and the ceiling it uses are machine-checked.
2. Both terms are readable on a drafter that already exists. The same condition replaces a transportation problem per head with two passes over the retained rows, certifying the split at vocabulary scale without training a second drafter, under controls that gate what may be read.
3. Marginal error does not determine what dependence can recover: at a fixed prefix the certificates separate 133 of 204 family pairs matched in estimated marginal error, and 120 of those gaps survive equalising the error itself, at a median certified lower bound of 0.108, while the preregistered comparison inside a band of width 0.05 gives 0.350 and carries sample composition with it. The target's own head does not behave like the families that concentrate mass.

4. A marginal-error calibration fitted on one prefix distribution does not transfer to another unverified: the one the code holdout arrives at by step 3000 sits below both prose sets, whose prefixes sit higher in first-position entropy, and the residual acquires structure in that entropy over training. Three seeded runs agree on every checkpoint verdict, though over half of that agreement is agreement that the checkpoint does not resolve.

2 BACKGROUND

Fix a prefix and let V be the vocabulary. Write P for the target's joint law over the next two positions, a distribution on $V \times V$, and Q for the drafter's. Write P_1, P_2 and q_1, q_2 for the first and second marginals of P and of Q. Write TV for total variation, half the L_1 distance. A distribution-preserving procedure that commits tokens proposed by Q induces a coupling of Q with P, and commits at most the mass that coupling places on agreement. That bound is attained and separates in two.

Theorem 1 (depth-two maximal-agreement coupling)**.** *Over all couplings of Q with P, the largest expected number of committed tokens, counting the one target token that follows the accepted prefix on every path, is*

$$3 - \mathrm{TV}(q_1, P_1) - \mathrm{TV}(Q, P), \tag{1}$$

and some coupling attains it.

Theorem 1 is the depth-two instance of the maximal-agreement coupling of Völlering (2016). In general the optimum over algorithms that verify a block while preserving the target distribution is smaller and is studied by Sun et al. (2025), but under $q_1 = P_1$ at depth two it is not: token-by-token verification (Leviathan et al., 2023) commits $3 - \mathrm{TV}(Q, P)$ in expectation on any such drafter, which Appendix A derives. So D_{rec} is the gap in expected committed length between the factorised drafter and the best one sharing its marginals, and not only the geometry an unrestricted coupling could exploit.

The first subtracted term charges the first position and the second charges the block, so a drafter whose first marginal is correct pays only the second. Changing q_2 moves the factorised law $q_1 \otimes q_2$ and the class of joint laws against which recoverable dependence is defined, so a comparison between drafters with different second marginals cannot say what dependence is worth on its own. The control that isolates it is the Fréchet class $\Gamma(q_1, q_2)$, the joint laws on $V \times V$ with marginals q_1 and q_2. Every member proposes the same tokens with the same frequencies at each position and differs only in how the two positions are arranged together. The factorised drafter is one member, and in equation 1 the first term is the same for all of them, so the gain from moving to another member Q' is the drop in joint total variation. Write $D_0 = \mathrm{TV}(q_1 \otimes q_2, P)$, the depth-two deficit, for what the factorised drafter loses, $M = \min_{Q' \in \Gamma(q_1, q_2)} \mathrm{TV}(Q', P)$ for what no rearrangement can remove, and $D_{\mathrm{rec}} = D_0 - M$ for what dependence can recover. Computing M in general is a transportation problem, given in Appendix A; for the architectures measured here it is not needed.

3 METHODOLOGY

3.1 DECOMPOSING THE DEFICIT

In the general case M is at least $\max\{\mathrm{TV}(q_1, P_1), \mathrm{TV}(q_2, P_2)\}$, and the difference, which is at least zero, is neither marginal error nor recoverable dependence: it is the cost of there being no joint law with those marginals near P. Denote it D_{comp}. Writing $\mu_1 = \mathrm{TV}(q_1, P_1)$ and $\mu_2 = \mathrm{TV}(q_2, P_2)$, the general accounting is

$$D_0 = \max\{\mu_1, \mu_2\} + D_{\mathrm{comp}} + D_{\mathrm{rec}}. \tag{2}$$

The two terms decompose one deficit and are not separately reducible: changing q_2 moves the class, so repairing the second marginal can raise D_0, which Appendix A shows. Both deployed head families remove the middle term by construction: they take the first drafted position from the target's own next-token head, so $q_1 = P_1$ exactly, which makes $\mu_1 = 0$ and, less obviously, $D_{\mathrm{comp}} = 0$.

Theorem 2 (the projection is a marginal distance)**.** *If $q_1 = P_1$ then $M = \mathrm{TV}(q_2, P_2) = \mu_2$, so the middle term of equation 2 vanishes and*

$$D_0 = \mu_2 + D_{\mathrm{rec}}. \tag{3}$$

Appendix A proves Theorem 2 and gives the general transportation formulation of M that the specialisation removes. The statement and its lemmas are machine-checked in Lean with mathlib, and a separate check runs both sides against that transportation problem in exact rational arithmetic, including the case where the hypothesis is removed, which permits the identity to break without requiring it.

3.2 THE RECOVERABLE TERM AND ITS CEILING

For tokens i and j in V, write $P(i, j) = P_1(i)P_{2|i}(j)$, so $P_{2|i}$ is the target's law at the second position after it has emitted i. Under $q_1 = P_1$, then, $D_0 = \mathbb{E}_{i\sim P_1}\mathrm{TV}(P_{2|i}, q_2)$ and $\mu_2 = \mathrm{TV}(\mathbb{E}_{i\sim P_1}P_{2|i}, q_2)$, so D_0 is a distance averaged over the first position and μ_2 is that same distance taken at the average. The difference,

$$D_{\mathrm{rec}} = \mathbb{E}_{i\sim P_1}\,\mathrm{TV}(P_{2|i},\, q_2) - \mathrm{TV}(P_2,\, q_2), \tag{4}$$

is the total variation lost in marginalising, and is non-negative because total variation is convex in its first argument. μ_2 is the part of the second head's error that survives averaging over the first position and D_{rec} is the part that does not.

Evaluating D_0 at $q_2 = P_2$ gives the target's own dependence $\tau_2 = \mathbb{E}_{i\sim P_1}\mathrm{TV}(P_{2|i}, P_2)$, equivalently $\mathrm{TV}(P_1 \otimes P_2, P)$. One triangle inequality per row gives $D_{\mathrm{rec}} \le \tau_2$, with equality at $q_2 = P_2$: a drafter cannot recover dependence the target does not have. For $\tau_2 > 0$, $\mathrm{reach} = D_{\mathrm{rec}}/\tau_2 \in [0, 1]$ is the share of the target's dependence still recoverable, the recoverable fraction the abstract names. Dividing by τ_2 removes the part of D_{rec} belonging to the target rather than the head, which makes it comparable between targets. When $\tau_2 = 0$ the positions are independent, nothing is recoverable, and reach is undefined.

3.3 THE CERTIFIED ESTIMATOR

Evaluating the decomposition exactly requires the target's joint law over $V \times V$: D_0, P_2 and τ_2 each depend on the conditional row $P_{2|i}$ for every candidate first token, and obtaining a row means evaluating the target once.

Choose $A \subset V$ and write $\delta = 1 - \sum_{i\in A} P_1(i)$ for the first-token mass left out. Take A to be the candidates of largest P_1 until δ falls below a budget of 2×10^{-3} or $|A|$ reaches a cap of 16384, whichever comes first; the fixed-target sweep ran looser, at 5×10^{-3} and 8192. A general estimator bounds M by a flow and a cut for any q_1 and needs several row-selection criteria; under $q_1 = P_1$ they collapse into the criterion above, which depends on P_1 and not q_2, and the specialised estimator built here takes that condition as given. The kept rows are exact over the whole vocabulary, so no column is truncated, and Appendix B gives the certified interval for D_{rec} that equation 3 implies. The cost is one target forward pass per retained row, 16041 for the whole of Medusa and 322470 for the other, against 32000 per position with no truncation; a further head at a measured position costs two passes over an array. The conditional rows $\{P_{2|i}\}_{i\in A}$ are computed once at a position and reused for every drafter compared there: every perturbation, every trained head, every seed and every checkpoint. That reuse costs memory, $|A| \times |V|$ doubles at a position and 2.6 GB at the largest row set here, and the whole Medusa measurement is 167 s on one A4000. How many rows the budget buys differs by a factor of four between prefix distributions, which Appendix E gives.

A subscript lo or hi on a quantity names the lower or the upper endpoint of its certified interval. The certificates cover the truncation, bounding what dropping rows can do to D_0, μ_2 and τ_2. They do not cover the bootstrap intervals, the rank correlations or the numerical implementation, which the reverse controls address separately. τ_2 is estimated with a certified overlap bound of its own, given in Appendix B. When $\tau_{2,\mathrm{lo}} > 0$ the certified interval for reach is $[D_{\mathrm{rec,lo}}/\tau_{2,\mathrm{hi}},\ D_{\mathrm{rec,hi}}/\tau_{2,\mathrm{lo}}]$, and otherwise the certificate says nothing about the ratio and the position carries no reach claim. Where a scalar is needed it is that ratio of midpoints; an interval on a mean residual is the cluster bootstrap rather than the certificate.

A position whose interval on D_{rec} is as wide as τ_2 itself cannot identify the ratio. Call a position informative when the width of the certified interval for D_{rec} over the midpoint of τ_2 is below 0.10, and read positions only there. The width is read once at the target's own second-position head, the native head, and reused for every drafter compared there, so no drafter's own certificate chooses the positions it is judged on; Appendix E gives the family this cannot be done for. That rule is about

Table 1: Per-position decisions at a threshold of 0.02 committed tokens, read from the certified interval on D_{rec} alone. Ruled out means the interval lies wholly below the threshold, opportunity that it lies wholly above, astride that it contains the threshold. The ceiling column counts the rule-outs that the proved bound $D_{\mathrm{rec}} \leq \tau_2$ already settles from the target alone; the split column counts those needing this head's own recoverable term. A rule-out is credited to the ceiling only where the τ_2 certificate is valid: 1 Medusa position returns an inverted one and is counted in the split column, where its own interval on D_{rec} settles it.

	positions	ruled out	ceiling	split	opportunity	astride
Medusa	128	58	49	9	67	3
multi-token prediction	128	13	7	6	114	1

identifying the ratio, and it is not the same test as numerical validity: an interval that comes back inverted is not a measurement of anything, and nothing below reads one, while a position that cannot identify the ratio can still bound the absolute term.

Three kinds of reverse control, given in Appendix C, decide whether a measured number may be read at all.

4 Experiments

Table 2 lists the chain, and Appendix E gives the protocols in full.

4.1 Two deployed head families

Medusa on Vicuna (Chiang et al., 2023) is dominated by its marginal error, $\mu_2 = 0.5003$ against $D_{\mathrm{rec}} = 0.0778$, and Meta's multi-token-prediction 7B by its dependence term, 0.1607 against 0.3332, both decomposed in Table 3. Nothing causal is read from the pair: architecture, base model, tuning and corpus all differ at once, and the multi-token-prediction checkpoint is a code model reading English. Read the way a drafter is read today, by marginal quality alone, the deficit is taken to be μ_2 and the dependence term is invisible. That reading misses 13% of Medusa's depth-two deficit and 67% of the other's. The rule of Section 3.3 keeps 67 and 108 of the positions, and on those that reading misses 21% and 70% of the deficit instead. Both regimes occur, and the cross-model confound sets up the next experiment. The decision the introduction opens with is about the absolute term rather than the ratio. Table 1 takes 0.02 committed tokens as the gain a rebuild would have to beat and reads each position's certificate against it, with 0 inverted intervals on D_{rec} on either family. Most of the rule-outs follow from the τ_2 ceiling alone, which needs nothing about the head; the split adds 9 of Medusa's 58 and 6 of the other's 13. Each entry is one prefix at one position, so a statement about a deployment needs a prefix distribution and its weights. The rule of Section 3.3 is the wrong filter for that question: it discards 57 of Medusa's 58 rule-outs, since a position can fail to identify the ratio and still bound the absolute term. The contrast reverses the preregistration, as Appendix E records.

4.2 Synthetic second marginals on a frozen target

The target is the multi-token-prediction 7B, and the prefixes and positions are held fixed: at every position P, P_1, the conditional rows, τ_2 and the Shannon entropy $H(P_1)$ of the target's own first-position law, in nats, are the same numbers for every drafter compared there, and only q_2 moves. The drafters are five one-parameter families anchored at the native second-position head, itself a q_2: one interpolates it toward the target's second marginal, which improves it, two redistribute its mass toward the uniform law and the first marginal, and two concentrate it by sharpening and truncation. The dispersing and concentrating families move the head in opposite directions at comparable marginal error. Two questions are fixed before the sweep is read: whether μ_2 predicts reach, and whether the five families agree at matched μ_2, as a threshold on their spread.

4.3 A SECOND TRAINED HEAD EVALUATED AGAINST THE SAME TARGET

The multi-token-prediction release ships a second checkpoint of the same architecture trained on fewer tokens, and that checkpoint supplies q_2 from its own trunk and head against the original target. Target, prefixes and positions are unchanged between the two drafters, over the 110 informative positions of 128, so P, P_1, τ_2 and $H(P_1)$ are identical numbers at every position and only the drafter differs. Its marginal error is comparable to that of the concentrating families over part of the sweep.

4.4 CHECKPOINTS ALONG THREE SEEDED TRAINING RUNS

We train the same future-token-head architecture on the frozen trunk of the same target, over an in-domain code corpus, with 3 independently seeded runs and five checkpoints at steps 10, 42, 173, 721, 3000, from a pilot.

4.5 PREFIX DISTRIBUTIONS AND THE ANALYSIS PROTOCOL

Three evaluation sets, in Table 4, differ in where the prefixes come from: the target's own frozen generations, 110 of 128 positions informative; a held-out split of the code corpus, in-domain by corpus type, 97 of 192; and abstracts from a fixed 2026 window, 173 of 192.

A position whose μ_2 falls outside the fitted curve's support receives no prediction, is counted, and is never clamped. That curve was frozen before this experiment was designed, no trajectory data enter it, and a control requires the re-expression to rebuild it under the original estimator.

The frozen calibration curve, a 14-knot linear interpolation fitted on the sweep, is the one the preregistration locked, read with the general estimator; the respecified curve is the same records, mask and bins re-expressed for the specialised estimator, before any trajectory is read. Write $f(\mu_2)$ for the curve in use and $e = \mathrm{reach} - f(\mu_2)$ for the residual. A positive residual is more recoverable dependence than the curve predicts at the same marginal error. The curve is a per-bin mean over the perturbations landing in each bin, so a statement that a head sits on it depends on that mixture; Appendix F checks one equal-family weighting, under which the between-set gaps move by less than the arms do. Both were fitted on the fixed-target sweep, so a residual read there is in-sample. Elsewhere the holdout has levels, which Appendix E gives. Every rank correlation below is Spearman's ρ.

5 RESULTS

Appendix E says which estimator and which curve each subsection below reports against.

5.1 MARGINAL ERROR AGAINST REACH

With the target, the prefixes and the conditional rows held fixed, a worse second marginal goes with a smaller reach. Over the informative positions of the sweep the rank correlation between μ_2 and reach is -0.918, across 375 records from 25 of 32 positions, one record being one position and one drafter. Figure 2 draws the 343 of those records inside the curve's support. One family runs toward the second marginal, the end at which theory forces reach to 1; without it the correlation is -0.820 over 275 records. The second question fixed before the sweep was read asked whether, inside a bin of μ_2 of width 0.05, the five families agree to within 0.10. The check fails in 9 of 10 bins, at a median spread of 0.224 and a largest of 0.350 with $[0.268,\ 0.462]$, and in the bootstrap that largest falls below the preregistered threshold with probability 0.000. A bin of width 0.05 is not a point, so part of that spread is the relation read at different places inside it: on residuals from the frozen curve the same statistic is 0.337, post hoc. A bin also mixes positions, which Appendix E counts, so the bin statistic carries prefix composition with the error geometry. Holding the position fixed removes it, and solving each family's parameter for one estimated μ_2 removes the binning: over 204 pairs at 24 positions the certified intervals for D_{rec} are disjoint on 133, at a median gap of 0.065 against the 0.017 the remaining mismatch in μ_2 is worth, so Proposition 3 converts 120 of them into proposals with exactly equal second-marginal error whose reach still differs, by a median certified 0.108, post hoc; Appendix E sets it out. Marginal error is a strong coordinate and not a sufficient one. Read

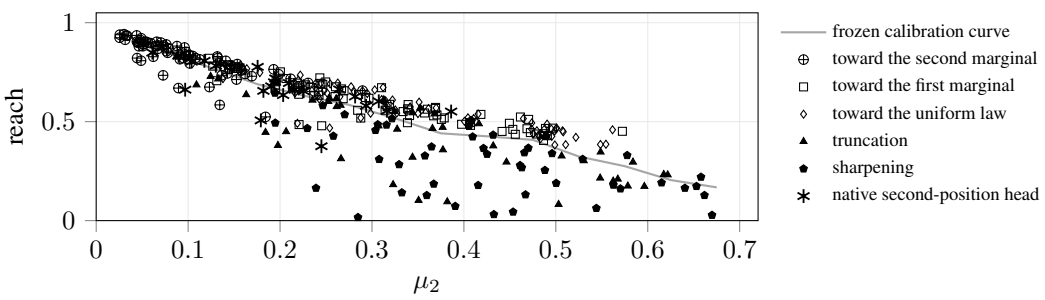


Figure 2: Reach against μ_2, one fixed target, only q_2 varying. Each mark is one position and one drafter; the asterisks are the native head. Among the perturbations, filled marks concentrate the head's mass and open ones spread it, and the crossed mark interpolates toward the target's own second marginal, which does neither. The clouds overlap, so the direction each family sits on is a statement about its mean, which Table 6 gives and Section 5.1 reads.

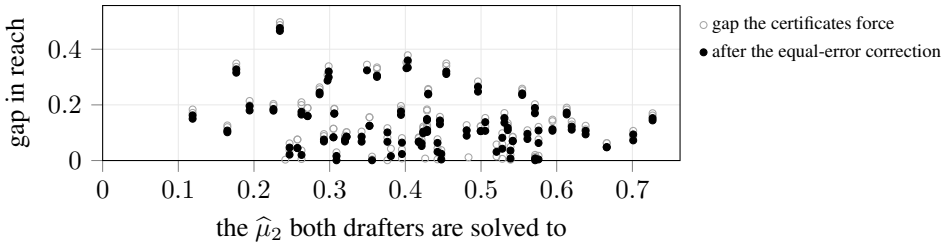


Figure 3: Every pair of drafters here sits at one estimated μ_2 at one position, solved for rather than binned, so the vertical distance between two of them is not the calibration curve read at two places. Rings are the gap the certified intervals force, on the 133 pairs of 204 whose intervals for D_{rec} are disjoint. Filled marks are what survives Proposition 3, which equalises the remaining difference in the true marginal error exactly and charges twice that difference against the gap.

at each deployed family's own μ_2, the coordinate reproduces that family's measured reach: 0.349 against Medusa's $[0.310,\ 0.389]$ and 0.714 against the other's $[0.706,\ 0.736]$, each inside its interval. What agrees is the mix of perturbations each window holds, not the relation alone; Appendix E gives the compositions that leave those intervals. The sweep's target is the multi-token-prediction model, so only the first leg transfers to a different family; the second transfers to different positions of the same one.

Table 6 says which way: at comparable marginal error the spreading families sit above the frozen calibration curve and the concentrating ones below, and interpolating toward the target's own second marginal lands closest.

5.2 A TRAINED HEAD AGAINST THE SYNTHETIC FAMILIES

A real head does not behave like the concentrating families. Held out against the sweep the two trained heads do not separate from the native one, which Appendix E reports; the separation appears when they are paired on one set of positions.

The sweep in Table 6 carries the native second-position head as one of its drafters, so it and the perturbations are read on one set of 25 positions. Paired by position it sits $+0.0604$ $[+0.0406,\ +0.0833]$ above truncation and $+0.1308$ $[+0.0920,\ +0.1728]$ above sharpening, both excluding zero, and -0.0028 $[-0.0112,\ +0.0059]$ from interpolation toward the second marginal, which does not.

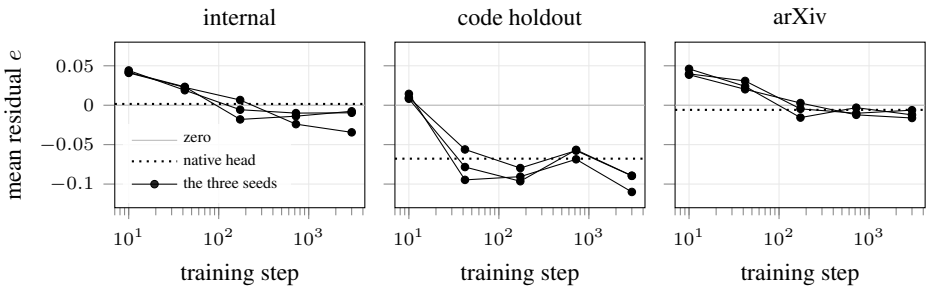


Figure 4: Mean residual along training, three seeded runs per panel as point estimates; the verdicts come from bootstrap intervals, summarised in Table 5. The two horizontal levels are zero and the target's native second-position head on that set, which coincide on both prose panels under the curve as fitted.

5.3 Seed-to-seed reproducibility within a distribution

Three seeded runs on a frozen representation are read at five checkpoints on each of three prefix distributions. The seeds agree on whether the mean residual excludes zero at 15 of the fifteen. That count is the weakest reading here: 8 of the fifteen agree on an interval that straddles zero. The continuous reading is stronger. At each set's worst checkpoint the seeds run 0.0267, 0.0385 and 0.0185 apart against excursions of 0.0595, 0.1070 and 0.0531 across the five. Appendix E gives which verdicts sit inside the bootstrap's own resolution and what fixing the positions does.

5.4 Calibration across the three prefix distributions

Figure 4 gives the trajectories. The two prose distributions start above the respecified curve and become indistinguishable from it. The native second-position head is indistinguishable from zero on both prose sets, on the curve as fitted; Appendix F gives a knot weighting under which the internal reading does not. The code trajectory is indistinguishable from zero at step 10, significantly negative by step 42, and stays there, at -0.1100 to -0.0896 by step 3000 against a native head of -0.0678. The calibration reached by step 3000 is prefix-distribution dependent, and the sets overlap in reach itself, 0.432 to 0.472: what separates them is where on the curve they sit. Low residuals occur more at low-entropy positions, and this design does not separate entropy from corpus composition or from where inside the support the positions sit. Restricting it to the entropy band the prose sets occupy puts all 3 across zero, the estimate moving while the width does not. On the direct step-3000 difference the code holdout sits below arXiv on all 3 seeds and below the internal set on 2, which Appendix F gives with the lower edge that undoes the first. The sets retain different numbers of informative positions, so a difference between two compares different subsets.

5.5 First-position entropy and the residual

The diagnostic fixed in advance is ρ between $H(P_1)$ and the residual e, in Figure 5. At step 10 the association is near zero, at most 0.073 in absolute value across the nine run and distribution combinations and negative on all three arXiv runs. By step 721 all nine are at least $+0.329$. The coupling appears with training, on one target and on positions selected using the same covariate, so what is established is its appearance, not that entropy causes the offset.

The coupling is not stable throughout training and does not grow at every step: 5 of the nine fall back between step 721 and step 3000, and Appendix E gives the spread at step 173.

6 Discussion

The recoverable term is the Jensen gap of equation 4, so it depends on how the head is wrong and not only on how wrong it is, and τ_2, which belongs to the target, bounds it, so a target whose positions are nearly independent leaves little to recover.

How the readings are organised, as an offset each prefix distribution carries plus an unspecified training-dependent term in $H(P_1)$, is set out in Appendix E, with nothing here fitted to it.

7 Related work

Work here differs in which feasible set is optimised over, which marginals are held fixed while that runs, and whether it reports an opportunity or one algorithm's gain. Völlering (2016) studies maximal agreement couplings, Ernst et al. (2018) constructs the complementary object, couplings that keep two different processes equal as long as possible, and Daliri et al. (2025) bounds what two parties can agree on with no shared randomness, all three characterising agreement at the coupling level. Leviathan et al. (2023) realises the single-position case of that coupling as an acceptance rule and reports end-to-end decoding speedups; Section 2 measures against the best coupling's value. Sun et al. (2025) optimises over procedures that verify a block while preserving the target distribution and shows it is best in that class; Weng et al. (2025) verifies a path through a token tree rather than token by token. Our measurements are of the opportunity at fixed marginals, not of an implementation's speedup.

A second line enlarges the proposal set, a direction orthogonal to what independence at fixed marginals costs. Candidates are organised into a token tree verified with tree attention (Miao et al., 2024), whose shape can be adapted (Wang et al., 2025); multi-draft selection is treated as optimal transport (Sun et al., 2023), solved by convex minimisation (Thomas & Pal, 2025), relaxed to a list (Rowan et al., 2025) or built by tree Monte Carlo (Hu & Huang, 2024); recursive rejection sampling is replaced to raise acceptance at later drafts (Sun et al., 2024); and maximal coupling is applied to Jacobi decoding (So et al., 2025). Lin et al. (2024) targets the cost of speculating, not latency.

Nearest is the work that builds or repairs that dependence. Ankner et al. (2024) makes each draft head a function of the tokens earlier heads sampled, and Li et al. (2024) drafts autoregressively in the target's second-to-top-layer feature space, feeding in the token sampled one step ahead to resolve which conditional a Medusa head predicts. Oda et al. (2026) analyses why factorised drafters degrade at later positions, Rusanovsky et al. (2026) corrects a block head whose tokens are individually plausible and jointly unlikely, Karimi et al. (2026) drafts dependent blocks, and Tang et al. (2026) decodes several tokens jointly. What none provides is the decomposition at fixed marginals: how much of a drafter's deficit comes from its marginals and how much from dependence.

8 Limitations

The split between the two prose distributions and the in-domain code holdout does not establish that code causes a negative offset: target, heads and evaluation set are all code, so the two distributions are nowhere separated factorially. Appendix F restricts the residual to the interior of the support, where the offset the trained heads acquire on the code set survives and the prose sets change.

We do not extrapolate to deeper blocks, where optimal agreement constrains more than two marginals, and a drafter whose first marginal comes from a learned head of its own, or from the target's head under a sampling transform the target does not share, falls under equation 2. The training evidence is about fitting a head to a fixed representation, not joint training. Under the hypothesis here D_{rec} is attainable in expected committed length, which Section 2 shows for token-by-token verification; what is not shown is that a proposal attaining it can be built, sampled and run cheaply.

9 Conclusion

When the first proposed position comes from the target's own next-token head, the depth-two deficit separates exactly into second-marginal error and a recoverable term, both readable on the drafter that already exists. What a rebuild could buy at the marginals that drafter has is then a measurement rather than a guess: on the positions the rule keeps it is 0.39 of a block's three committed tokens for the multi-token-prediction model and 0.14 for Medusa. That is an opportunity, not a speedup. Marginal error predicts what dependence can recover without determining it, and the offset that does not carry across tracks a deployment's prefix entropy.

Reproducibility Statement

Every experimental number in this paper is a macro generated from committed records by a single script, and the records it reads are pinned by SHA-256 in a manifest, `exp/RECORDS.sha256`, that `sha256sum -c` checks without running that script. The Lean development machine-checks 6 exported statements, all with no `sorryAx`, and ships as a package that builds. Appendix A gives the proofs and the general formulation, Appendix B the truncation budget and the conditions under which the estimator is certified, Appendix C the reverse controls, Appendix D the experiment chain and the tables behind the deployed families, the evaluation sets and the perturbations, Appendix E the training, row-selection and bootstrap protocols with the seeds they use, the populations each statistic is read on and the compositions behind the transfer legs, and Appendix F the post hoc support-edge probe. The supplementary material contains the records, the generator, the Lean development, the preregistrations and round results this paper cites, and `reproduce.sh`, which rebuilds every number and figure in the paper from them and compares the result against the committed copy the paper was built from.

Ethics Statement

This work involves no human subjects and no personal data. The measurements are made on publicly released model checkpoints run on rented hardware, given text prefixes drawn from a held-out split of a public Python corpus, arXiv abstracts from a fixed window, and the target model's own generations, all three described in Section 4. No data was collected from people for this study and none was released by it. The object of study is a property of decoding systems that are already deployed rather than a new capability, and the practical consequence of reading it wrongly is engineering effort spent on the repair that was not needed, which is what the reverse controls in Appendix C are for.

AI Use Statement

Use of Generative AI. The authors used ChatGPT by OpenAI and Claude by Anthropic to assist with language editing, clarity, concision, and organization of author written text. All technical content, claims, experimental results, analyses, and citations were produced and verified by the authors. All AI assisted edits were reviewed and revised by the authors, who take full responsibility for the content of the paper.

References

Zachary Ankner, Rishab Parthasarathy, Aniruddha Nrusimha, Christopher Rinard, Jonathan Ragan-Kelley, and William Brandon. Hydra: Sequentially-dependent draft heads for Medusa decoding. *CoRR*, abs/2402.05109, 2024. doi: 10.48550/arXiv.2402.05109. URL https://doi.org/10.48550/arXiv.2402.05109.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235. PMLR / OpenReview.net, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/, March 2023.

Majid Daliri, Christopher Musco, and Ananda Theertha Suresh. Coupling without communication and drafter-invariant speculative decoding. In *IEEE International Symposium on Information Theory, ISIT 2025, Ann Arbor, MI, USA, June 22-27, 2025*, pp. 1–6. IEEE, 2025. doi: 10.1109/ISIT63088.2025.11195663. URL https://doi.org/10.1109/ISIT63088.2025.11195663.

DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/arXiv.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/arXiv.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

P. A. Ernst, W. S. Kendall, G. O. Roberts, and J. S. Rosenthal. Mexit: Maximal un-coupling times for stochastic processes, 2018. URL https://arxiv.org/abs/1702.03917.

GLM-4-5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *CoRR*, abs/2508.06471, 2025. doi: 10.48550/arXiv.2508.06471. URL https://doi.org/10.48550/arXiv.2508.06471.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15706–15734. PMLR / OpenReview.net, 2024.

Zhengmian Hu and Heng Huang. Accelerated speculative sampling based on tree Monte Carlo. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19216–19251. PMLR / OpenReview.net, 2024.

Amirmohammad Karimi, Chao Gao, and Negar Hassanpour. Dblast: Dependent block drafting for stochastic speculative decoding, 2026. URL https://arxiv.org/abs/2608.05448.

Kimi Team. Kimi K3: Open frontier intelligence. *CoRR*, abs/2607.24653, 2026. doi: 10.48550/arXiv.2607.24653. URL https://doi.org/10.48550/arXiv.2607.24653.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 2023.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28935–28948. PMLR / OpenReview.net, 2024.

Chi-Heng Lin, Shikhar Tuli, James Seale Smith, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Slim: Speculative decoding with hypothesis reduction. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, volume NAACL 2024 of *Findings of ACL*, pp. 1005–1017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.63. URL https://doi.org/10.18653/v1/2024.findings-naacl.63.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024*, pp. 932–949. ACM, 2024. doi: 10.1145/3620666.3651335. URL https://doi.org/10.1145/3620666.3651335.

Yuma Oda, Ryan Mathieu, Roman Knyazhitskiy, and Artur Chakhvadze. Trees from marginals: Autoregressive drafting with factorized priors. *CoRR*, abs/2607.06763, 2026. doi: 10.48550/arXiv.2607.06763. URL https://doi.org/10.48550/arXiv.2607.06763.

Joseph Rowan, Buu Phan, and Ashish Khisti. List-level distribution coupling with applications to speculative decoding and lossy compression. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025,*

San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025, 2025.

Matan Rusanovsky, Yoav Miron, Roy Uziel, Omer Belhasin, Ran Zilberstein, Maor Ashkenazi, and Michael Elad. Lilicorr: Lightweight likelihood correlation of parallel drafts for speculative decoding, 2026. URL https://arxiv.org/abs/2608.20530.

Junhyuk So, Hyunho Kook, Chaeyeon Jang, and Eunhyeok Park. MC-SJD : Maximal coupling speculative Jacobi decoding for autoregressive visual generation acceleration. *CoRR*, abs/2510.24211, 2025. doi: 10.48550/arXiv.2510.24211. URL https://doi.org/10.48550/arXiv.2510.24211.

Ryan Sun, Tianyi Zhou, Xun Chen, and Lichao Sun. Spechub: Provable acceleration to multi-draft speculative decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 20620–20641. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1148. URL https://doi.org/10.18653/v1/2024.emnlp-main.1148.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix X. Yu. Spectr: Fast speculative decoding via optimal transport. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Ziteng Sun, Uri Mendlovic, Yaniv Leviathan, Asaf Aharoni, Jae Hun Ro, Ahmad Beirami, and Ananda Theertha Suresh. Block verification accelerates speculative decoding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Zhiwei Tang, Yuanyu He, Yizheng Han, Wangbo Zhao, Jiasheng Tang, Fan Wang, and Bohan Zhuang. K-forcing: Joint next-k-token decoding via push-forward language modeling. *CoRR*, abs/2606.10820, 2026. doi: 10.48550/arXiv.2606.10820. URL https://doi.org/10.48550/arXiv.2606.10820.

Rahul Krishna Thomas and Arka Pal. Global resolution: Optimal multi-draft speculative sampling via convex minimization. *CoRR*, abs/2511.15898, 2025. doi: 10.48550/arXiv.2511.15898. URL https://doi.org/10.48550/arXiv.2511.15898.

Florian Völlering. On maximal agreement couplings, 2016. URL https://arxiv.org/abs/1608.01511.

Jikai Wang, Yi Su, Juntao Li, Qingrong Xia, Zi Ye, Xinyu Duan, Zhefeng Wang, and Min Zhang. Opt-tree: Speculative decoding with adaptive draft tree structure. *Trans. Assoc. Comput. Linguistics*, 13:188–199, 2025. doi: 10.1162/tacl_a_00735. URL https://doi.org/10.1162/tacl_a_00735.

Yepeng Weng, Qiao Hu, Xujie Chen, Li Liu, Dianwen Mei, Huishi Qiu, Jiang Tian, and Zhongchao Shi. Traversal verification for speculative tree decoding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

A Proofs and the general formulation

Write $Q(i, j) = P_1(i)r_i(j)$ for a drafter whose first marginal is the target's. Token-by-token verification accepts the first position always, and given a first token i accepts the second with probability $\sum_j \min\{r_i(j), P_{2|i}(j)\}$; averaging over $i \sim P_1$ gives $1 - \mathrm{TV}(Q, P)$, so the expected committed length is $3 - \mathrm{TV}(Q, P)$, which is the value of equation 1. The factorised drafter commits $3 - D_0$ and the best member of $\Gamma(P_1, q_2)$ commits $3 - \mu_2$, their difference being D_{rec}. Nothing here says such a drafter is cheap to construct, to sample from, or faster in wall-clock time.

Proof of Theorem 2. The lower bound $M \geq \mu_2$ holds for any q_1, because total variation does not increase under marginalisation. The upper bound needs a member of the class that attains it, and the hypothesis makes one easy to exhibit. Scale each column j of P down so that it carries its shared mass $\min\{P_2(j), q_2(j)\}$. That leaves surplus mass in each row. Redistribute the surplus of row i across exactly those columns with $q_2(j) > P_2(j)$, in proportion to their deficits $q_2(j) - P_2(j)$. The row marginals of P are already the required ones, so mass moves inside each row and no feasibility condition has to be checked. The result sits pointwise below P on the columns that were scaled down and pointwise above it on the columns that received surplus, so its overlap with P is exactly $1 - \mu_2$. □

Computing M without the hypothesis is a transportation problem. Total variation is one minus overlap, so $M = 1 - F^*$ where F^* is the largest mass of P that can be kept while meeting the two marginals, and that is a maximum flow on the network with capacity $q_1(i)$ from the source to row i, $P(i, j)$ from row i to column j, and $q_2(j)$ from column j to the sink. The dual gives an elementary lower bound from any rectangle: for every $Q' \in \Gamma(q_1, q_2)$ and every pair of subsets S and T of V, a Bonferroni argument puts at least $q_1(S) + q_2(T) - 1$ of Q' on $S \times T$, so

$$\mathrm{TV}(Q', P) \geq q_1(S) + q_2(T) - 1 - P(S \times T),$$

and the largest such rectangle attains M, one minus the minimum cut. The two degenerate rectangles, $T = V$ and $S = V$, recover the two marginal total variations, which is the bound the specialisation replaces with an equality.

Theorem 2 and the lemmas it rests on are machine-checked in Lean `4.34.0` against mathlib at revision `5ed29652`. The development is 1339 lines and 6 of 6 exported statements depend only on `propext`, `Classical.choice` and `Quot.sound`. So do 6 further statements the same package proves and this paper otherwise states without proof: Theorem 1 with its attaining coupling constructed; $D_{\mathrm{rec}} = D_0 - M$ for any M attained in the class, which is equation 2 once D_{comp} is named; a ceiling that holds for an arbitrary q_1, $D_{\mathrm{rec}} \leq \mathrm{TV}(q_1 \otimes P_2, P)$, which reduces to τ_2 when $q_1 = P_1$ and need not equal it otherwise, so the τ_2 form does not hold without the hypothesis: at an independent target with $q_1 \neq P_1$ the two marginal errors alone leave room to rearrange, giving $D_{\mathrm{rec}} > 0 = \tau_2$; the rectangle bound above; and two on a pair of targets agreeing in both marginals, that no law is within $1/2$ of both and that one matching the first marginal commits at most $5/2$ on one of them.

At an almost deterministic target with $q_1 = P_1$, a head with $\mu_2 = 0.10$ loses $D_0 = 0.10$ and commits 2.90 of three tokens, while the head whose second marginal is exact loses 0.18 and commits 2.82. Ranking the two repairs is a different question from bounding either. A separate check runs both sides against the transportation problem in exact rational arithmetic, over 240 instances drawn from 4 families, dense Dirichlet draws and three chosen to break estimators: sparse supports, deterministic conditionals, and a q_2 supported where P_2 is not. The worst deviation is 0. Turning to the measured records, since $q_1 = P_1$ the covariate $H(P_1)$ belongs to the target and the prefix, so it is identical for every drafter compared at a position and does not move when a head is trained, so the entropy readings are not a property the training changes. The hypothesis itself is architectural and is asserted rather than measured. The first drafted position is read from the target's own next-token head, so each measurement passes one array as both q_1 and P_1, and μ_1 is zero at every record by construction rather than by agreement. What the records check is the consequence. Over 1136 records, of which 512 measure a drafter from a released checkpoint and the rest are the synthetic sweep and two rehearsals, meaning runs of the same measurement at a smaller row cap and on fewer positions, kept because they carry the same fields, the cut lower bound on M never reaches the certified upper end of μ_2, coming closest at 1.6×10^{-5} in one of the rehearsals, 1.9×10^{-5} in the fixed-target sweep the curves are fitted on, and no closer than 7.7×10^{-4} over the three deployed-family files, so the certified lower bound on D_{comp} stays at zero everywhere. The resolution of that check is the median distance between the two, 3.5×10^{-3}, and a compatibility term below it would not be visible. Most of that distance is not the budget: the μ_2 certificate's median half-width is 1.0×10^{-3} and beside it 2.5×10^{-3} is how far the rectangle search leaves the cut bound below the certified midpoint of μ_2, which carries no certificate of its own. The same check is run with the hypothesis removed, which permits the identity to break without requiring it, and over 200 instances with $q_1 \neq P_1$ it does so in 145. The witness above is also built explicitly and its two marginals verified, over 160 instances.

Proposition 3 (equal-error witness)**.** *Fix a prefix, let $q_1 = P_1$, and let q and q' be second-position proposals with $\mu_2(q) \geq \mu_2(q')$ and $\mu_2(q) > 0$. Write $d = \mu_2(q) - \mu_2(q')$ and*

$$q^\star = \frac{\mu_2(q')}{\mu_2(q)}\, q + \Big(1 - \frac{\mu_2(q')}{\mu_2(q)}\Big) P_2. \tag{5}$$

Then $\mu_2(q^\star) = \mu_2(q')$ exactly and $|D_{\mathrm{rec}}(q^\star) - D_{\mathrm{rec}}(q)| \leq 2d$, so if the certified intervals for $D_{\mathrm{rec}}(q)$ and $D_{\mathrm{rec}}(q')$ are disjoint with a gap $G > 2d$, two proposals exist at that prefix whose second-marginal errors are equal and whose recoverable terms differ by at least $G - 2d$.

Proof. Write $\lambda = \mu_2(q')/\mu_2(q) \in [0, 1]$. Then $q^\star - P_2 = \lambda(q - P_2)$, so $\mathrm{TV}(q^\star, P_2) = \lambda\mu_2(q) = \mu_2(q')$, and $q^\star - q = (1 - \lambda)(P_2 - q)$, so $\mathrm{TV}(q^\star, q) = (1 - \lambda)\mu_2(q) = d$. For any two proposals $\mathrm{TV}(P_{2|i}, \cdot)$ and $\mathrm{TV}(\cdot, P_2)$ each move by at most TV between them, so by equation 3 $|D_{\mathrm{rec}}(q) - D_{\mathrm{rec}}(\tilde{q})| \leq 2\,\mathrm{TV}(q, \tilde{q})$. Applying this to q and $q^\star$ gives the bound, and the last claim follows because q' is unchanged. □

The construction needs P_2, which the estimator does not have, so it is not a drafter this paper builds; what it certifies is that one exists. It also leaves the family: $q^\star$ need not lie on the path the perturbation followed, so the statement it supports is that marginal error does not determine the recoverable term, not that two members of these families sit at one μ_2.

B THE TRUNCATION BUDGET

Write $\widehat{D}_0 = \sum_{i \in A} P_1(i)\mathrm{TV}(P_{2|i}, q_2)$, $\widehat{P}_2 = \sum_{i \in A} P_1(i) P_{2|i}$ and $\widehat{\mu}_2 = \mathrm{TV}(\widehat{P}_2, q_2)$, where $\widehat{P}_2$ carries mass $1 - \delta$ rather than one and total variation is read as half the L_1 distance, which is defined for it. The dropped rows contribute between 0 and δ to D_0, and $\widehat{P}_2$ differs from P_2 by δ in total mass, so

$$D_0 \in [\widehat{D}_0,\ \widehat{D}_0 + \delta], \qquad \mu_2 \in \big[\max(0, \widehat{\mu}_2 - \delta/2),\ \min(1, \widehat{\mu}_2 + \delta/2)\big], \tag{6}$$

and equation 3 turns equation 6 into a certified interval for the recoverable term. Writing $D_{0,\mathrm{lo}}$ and $D_{0,\mathrm{hi}}$ for the endpoints of the first interval in equation 6 and $\mu_{2,\mathrm{lo}}$ and $\mu_{2,\mathrm{hi}}$ for those of the second,

$$D_{\mathrm{rec}} \in \big[\max(0,\ D_{0,\mathrm{lo}} - \mu_{2,\mathrm{hi}}),\ D_{0,\mathrm{hi}} - \mu_{2,\mathrm{lo}}\big]. \tag{7}$$

For an arbitrary q_1 the corresponding bound on M is a transportation problem and each candidate q_2 needs its own solve. Under $q_1 = P_1$, $M = \mu_2$ exactly, so once the conditional rows exist each further head costs two passes over an $|A| \times |V|$ array.

On the internal set at the native second-position head the specialised interval on D_{rec} has mean width 2.66 times smaller than the general route's and lies inside it at all 128 positions. The same rows give $\widehat{O}$, the overlap between the retained joint and the product of its marginals, from which τ_2 is read. Writing $P_2 = \widehat{P}_2 + R$ with $R \geq 0$ of mass δ, the dropped rows remove at most δ of that overlap and replacing P_2 by $\widehat{P}_2$ moves it by at most δ again, so τ_2 lies in $[1 - \widehat{O} - 2\delta,\ 1 - \widehat{O} + \delta]$, wider below than above because one substitution can only remove overlap while the other moves it either way.

C REVERSE CONTROLS

The certificate places D_{rec} above its τ_2 ceiling, in the sense that the lower end of one exceeds the upper end of the other, at 1 of Medusa's positions and 0 of the other family's. The intervals are disjoint far more widely, at 87 and 121 of them, in the direction $D_{\mathrm{rec}} \ll \tau_2$ that the ceiling permits, in every case but that one crossing. On the positions the informativeness rule keeps they are disjoint at all 67 and all 108, with 0 and 0 above the ceiling, so every position where the two intervals meet is one the rule removes. At that position the retained row set holds a single row and the certificate for τ_2 comes back inverted, its upper end below zero by about 10^{-8}, which is what puts D_{rec} above it; every rule that reports a ratio drops the position, and so does every quantity here that divides by τ_2: the mean τ_2 of Table 3, its pooled reach and that reach's certified interval are read on the 127 of 128 Medusa positions whose τ_2 certificate is valid, because a sum of bounds that includes an inverted bound is not bounded by the derivation that produced the others. Over every position the pooled

reach is 0.3484 against 0.3484, which measures sensitivity and does not repair the derivation. The absolute reading of Section 4.1 keeps the position, since it does not divide by τ_2 and that position's interval on D_{rec} is ordered, and the ceiling column of Table 1 does not credit it, since a rule-out attributed to the target's own dependence needs a τ_2 certificate that holds. The two families are not equally well resolved: the median retained row count is 4 against 1893, and 34 of Medusa's positions put the midpoint of D_{rec} above that of τ_2 against 0 on the other, which is what too few rows produces. The share a marginal-only reading misses, 13% and 67%, comes from applying the field's own criterion to these positions, and no prior paper reports it. The bound on the settling offset is the widest prose interval at step 3000, -0.0790.

Three kinds of reverse control decide whether a measured number may be read. Each is a gate: a failure invalidates the downstream measurement instead of widening its error bar. Two of the controls below are on their second version, and in both cases the first version failed and the failure changed the paper, so each is given with the version it replaced.

The first is a theoretical anchor. At $q_2 = P_2$ the theory forces $\mu_2 = 0$, $D_{\mathrm{rec}} = \tau_2$ and reach exactly 1. The sweep approaches that end by interpolating to $\widehat{P}_2$ scaled to mass one, so μ_2 there is zero only up to the dropped mass, and the control asks that μ_2 stay within that budget and that 1 lie inside the certified interval for reach. As first written it asked the midpoint itself to land within one per cent of 1, which is a demand on resolution and not on validity. It failed. The informativeness rule of Section 3.3 was written in the same response, so the threshold that defines every population in this paper was set by a control that failed.

The second is a cross-check against the general flow and cut estimator, which is the validated instrument. The specialised interval has to intersect it, over 240 synthetic positions in 6 regimes and again on real positions, with 0 disagreements. On the synthetic positions the result is stronger than intersection: the specialised interval lies strictly inside the general one at all 240 of them. The real-position half ran on the internal set alone, 128 positions, and not on the code holdout or on arXiv, where the general route was not run beside the specialised one. That is a cost decision and not an impossibility, and it leaves the two external sets carrying this control from the synthetic positions and the internal set only.

The third kind covers the execution path, in four checks: the head index each output belongs to, the reuse of cached target rows against a recomputation, the trunk control, which compares the re-implemented shared trunk against the official forward, and the first-token law recomputed across pipeline stages, which agrees to the bit over 208 items across all three evaluation sets with 0 mismatched. The cached-row one reports a magnitude rather than a bit identity, and it licenses reading D_{rec} at all: the cached path and a plain recomputation differ on D_{rec} itself by 2.5×10^{-5} on Medusa and 1.7×10^{-4} on the multi-token-prediction family, measured at three positions per family at a cap of 64 retained rows and then stamped into every record, on the deployed families and not on the three evaluation sets.

That trunk control is the strongest of those four and the one the training rests on: at matched batch the re-implemented trunk reproduces the official forward bit for bit, with a largest absolute difference of 0. It is also on its second version. The first compared logits after the head across batch sizes and required identical top-1 at every position, and it returned a failure; the diagnostic that replaced it found the official model disagreeing with itself across batch sizes at the same rate, so the comparison was moved to matched batch, where the states the training consumed are the states the gate reads. The head built from those states is checked separately over 4096 positions, where it makes 0 top-1 disagreements and differs from the official head by at most 0.0072 in total variation.

Over 240 exact instances the two inequalities the decomposition proves, $D_{\mathrm{rec}} \le \tau_2$ and $D_{\mathrm{rec}} \ge 0$, are violated 0 and 0 times, and the ceiling is attained with equality at 49 of them, so the control exercises the boundary rather than only the interior.

D SUPPLEMENTARY TABLES

Table 2: Each experiment, what it varies, what it holds fixed, and the confound it removes from the row above.

experiment	what varies	what is held fixed	confound removed
deployed-family contrast	model family	no causal control	none, motivation only
fixed-target intervention	q_2, size and shape	target, prefixes	cross-model
trained-head holdout	q_2, a second real head	target, prefixes	synthetic versus trained
seeded trajectories	training state	trunk, protocol	two checkpoints coinciding
prefix distributions	the prefix source	the trained heads	one prefix distribution

Table 3: The decomposition on two deployed parallel-head families. Every column of a row is read on the same positions, the 127 and 128 whose τ_2 certificate is valid, so the reach column is the D_{rec} column over the τ_2 column; over all 128 positions the first two columns are 0.4999 and 0.0772 for Medusa and 0.1607 and 0.3332 for the other. The positions are 8 positions from each of 16 prompts written for this work, continued by the target itself, the same prompts for both families. The list is `PROMPTS` in `measure/t1_medusa.py` of the supplement. Reach is pooled as $\sum D_{\mathrm{rec}} / \sum \tau_2$ because τ_2 falls below 10^{-3} at 30 of the Medusa positions, all 30 of them among the 34 midpoint crossings of Section 4.1, where reach is not identifiable and a mean of per-position ratios is numerically ill-conditioned. Its certified interval, pooled the same way, is $[0.331,\ 0.366]$ for Medusa and $[0.705,\ 0.735]$ for the other.

family	μ_2	D_{rec}	τ_2	reach
Medusa on Vicuna	0.5003	0.0778	0.2234	0.3484
Meta multi-token-prediction 7B	0.1607	0.3332	0.4628	0.7198

Table 4: Where each evaluation set's prefixes come from, and how many of its positions are informative. The internal count comes from the committed general-estimator certificate, kept frozen; the other two from this round's specialised certificate, as Appendix E sets out.

set	prefix source	positions	informative
internal	the target's own frozen generations, English prose	128	110
code	held-out split of the code corpus, same corpus type	192	97
arXiv	abstracts from a fixed 2026 window, scientific prose	192	173

Table 5: Whether the mean residual excludes zero, under the respecified calibration. Each checkpoint entry is three seeded runs agreeing; the native head column is that one fixed head. pos and neg mean the mean residual excludes zero on that side, 0 that it straddles zero.

	step 10	step 42	step 173	step 721	step 3000	native head
internal	pos	0	0	0	0	0
code	0	neg	neg	neg	neg	neg
arXiv	pos	pos	0	0	0	0

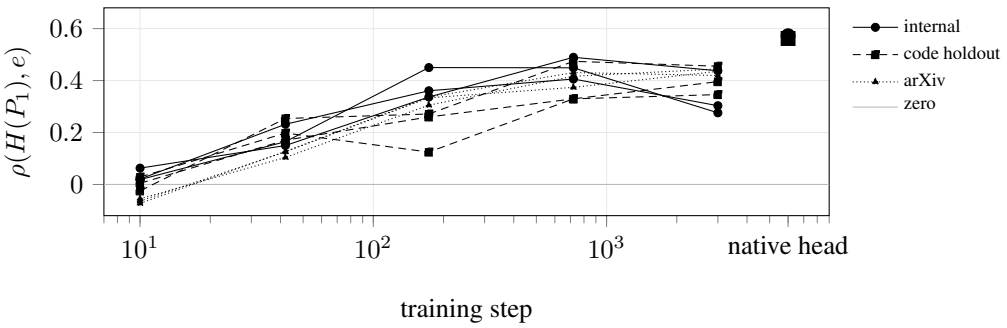


Figure 5: Rank correlation between $H(P_1)$ and the residual against training step, three seeds per distribution. The larger markers are the target's native second-position head read on each of the three sets, and they nearly coincide. That head is not a checkpoint, and the tick it sits on carries its name rather than a step.

Table 6: The five perturbation families on the sweep's own scale, residuals against the frozen calibration curve. kept and out count the family's informative records inside and outside the curve's support. native head is the native second-position head. Intervals are cluster bootstraps over prompts.

family	direction	kept	out	mean e	sd e	95% CI
toward the second marginal	improves	72	28	$+0.0068$	0.0538	$[-0.0124, +0.0225]$
toward the first marginal	spreads	75	0	$+0.0500$	0.0499	$[+0.0369, +0.0634]$
toward the uniform law	spreads	75	0	$+0.0608$	0.0504	$[+0.0454, +0.0739]$
truncation	concentrates	47	3	-0.0632	0.1210	$[-0.1039, -0.0260]$
sharpening	concentrates	49	1	-0.1323	0.1447	$[-0.1761, -0.0841]$
native head	reference	25	0	$+0.0020$	0.0806	$[-0.0223, +0.0231]$

E PROTOCOLS

The target is checkpoint `7B_1T_4` of `facebook/multi-token-prediction` (Gloeckle et al., 2024), 32 layers of width 4096 over a vocabulary of 32000, with 4 future-token heads. The held-out trained head of Section 5.2 is `7B_200B_4`, the sibling checkpoint in that release trained on fewer tokens. The deployed-family contrast reads the target and `FasterDecoding/medusa-vicuna-7b-v1.3`, the released Medusa heads for `lmsys/vicuna-7b-v1.3`. The target proposes 4 positions; everything here reads the first two, and the target's second-position head is the one called native throughout.

The code corpus is codeparrot/codeparrot-clean-valid, file file-000000000054.json.gz, whose SHA-256 begins `7f8c20a737c90847`. The trunk cache holds 4.0M tokens of trunk state, 7812 contiguous chunks of 512 tokens taken from inside single records of that corpus, cut within-record, no packing, tail dropped, with no chunk crossing a record boundary. Each head is the target's own extra-heads block, so the only difference from the native second-position head is training and not design, and the trunk contributes only the cached states. It is trained with cross-entropy on the next-position token, read through the target's own final norm and output embedding, with batch 4 and learning rate 10^{-4} under AdamW at zero weight decay. Nothing outside the head is updated. The chunk order is re-drawn at every epoch boundary from the same seeded generator that fixes the initialisation. The five checkpoint steps were chosen from a short pilot run whose numbers appear nowhere in the results and were committed before any of the three runs started.

Heads and positions are new to every reading outside the sweep, but prompts are not: the internal set reuses 13 of its 16 from the sweep the curve was fitted on and brings 3 of its own, while the code and arXiv sets are new prompts throughout. The internal evaluation set is the target's own continuation of the same 16 prompts, 8 positions read from each, sampled once and then frozen. They carry the same prompts and the same position indices as the multi-token-prediction row of Table 3, and both are read with the general estimator under the same budget and cap, but they are not the same

measurement. The internal set reads the sequences the sweep froze; the deployed row generated its own continuations, and a sampling draw under half precision diverged at 1 of the 16 prompts, so 7 of the 128 positions hold different tokens. On the 121 that reproduce each other both keep the same 103 positions, so the whole of the difference between 110 and 108 sits in the divergence and none of it in the rule. The retained row counts differ for the same reason. The prompts are short instructions written for this work rather than drawn from a corpus, so the internal set is the only one whose prefixes have no external provenance. Every distribution read anywhere in the paper, in that sampling and in all measurement, is the model's softmax over the full vocabulary at temperature one, with no truncation and no token masked. The code holdout is 96 records with 2 positions from each. A record is held out when the first four bytes of the SHA-256 of its content hash, read as one big-endian integer, are zero modulo sixteen, so the split is a property of the record and not of an ordering, and the trunk cache is built from the train side of that same split. The arXiv set is the 96 abstracts that remain after 1 was excluded for being too short, submitted between 2026-08-20 and 2026-09-17, drawn under seed 20260919 from a pool of 1160 formed by taking the same number of each day's earliest submissions on every day of the window, with 2 positions taken from each. No model-derived quantity enters that selection.

The covariate $H(P_1)$ belongs to the target and the prefix, but the coupling of Section 5.5 is a correlation with a residual, so it depends on the head and on the curve as well, and it differs between targets: read on the kept positions of Medusa's that also fall inside the curve's support it is $+0.114$ over 45 of the 67 the rule keeps, against $+0.573$ over all 108 of the multi-token-prediction family's, whose prefixes span a far wider range of $H(P_1)$. The two deployed families also differ in the covariate of Section 5.4: median $H(P_1)$ on the positions the rule keeps is 1.01 for Medusa against 2.90 for the multi-token-prediction model, which is the separation the code holdout and the two prose sets show, so the opposite regimes of Section 4.1 and the offset of Section 5.4 are ordered by the same quantity. The populations of the three sets are not selected by the same instrument. The internal mask is the committed general-estimator certificate, kept frozen so that an improvement to the estimator cannot move a population already reported against; the code and arXiv sets have no committed history, so theirs come from this round's specialised certificate. The specialised rule would keep 117 internal positions rather than 110, and on those 1 of the three seeds at step 3000 excludes zero, so the agreement count in Section 5.3 is one lower on that population. At 10 of the sweep's 480 records the midpoint of D_{rec} also sits above the midpoint of τ_2, all of them drafters at one position whose retained row set holds a single row, and none of them survives the informativeness rule. The fixed-target sweep runs on the same prefixes as the deployed-family measurement, the 16 written for this work, continued by the target itself. Its 32 prefix and position pairs come from 13 of them, and 25 of those pairs are informative. The replication round preregistered that marginal error would dominate the recoverable term on the multi-token-prediction family by a factor above two, as it does on Medusa at 6.43. It came back at 0.48, reversed, and the opposite-regime contrast the paper opens with is that reversal rather than a designed comparison. All 3 runs use seeds 1, 2, 3, and every cluster bootstrap in the paper is drawn under seed 20260920.

The perturbation families are built at each position from the native second-position head q. Three of them interpolate, taking $(1-\lambda)q + \lambda X$ normalised, with X the retained-row second marginal at $\lambda \in \{0.25, 0.5, 0.75, 1\}$, the uniform law at $\lambda \in \{0.1, 0.25, 0.5\}$ and P_1 at $\lambda \in \{0.1, 0.25, 0.5\}$. Sharpening takes $q^{1/T}$ normalised at $T \in \{0.5, 0.7\}$, and truncation keeps the k largest entries and scales them to mass one, at $k \in \{8, 64\}$. Counting the native head, that is 15 drafters at every position.

Matching two families at one μ_2 rather than inside a bin takes a parameter solve. Interpolation toward P_2 satisfies $\mathrm{TV}((1-\lambda)q + \lambda P_2, P_2) = (1-\lambda)\mathrm{TV}(q, P_2)$, so it moves μ_2 down only, and truncation is indexed by an integer, which leaves the two interpolations away from P_2 and sharpening. Each is reparameterised to run from the native head to the end of its family and bisected for the μ_2 a level asks for, at 3 levels placed inside the interval all three reach, between 0.12 and 0.73 against a median native μ_2 of 0.19. That run selects the sweep's own positions and measures the native head a second time, where it differs from the sweep by 0 at all 32 of them. The solve has no level at 1 informative position, because fewer than two families move μ_2 above the native head there, and 5 family and position pairs move it down instead. What the solve matches is the estimated $\widehat{\mu}_2$ of Appendix B, not μ_2, and the two differ by at most the dropped first-token mass δ, which is 8.4×10^{-3} at worst here. The dropped rows are a mixture both drafters share, which does not cancel in the difference: a column where one head sits above the retained marginal and the other below

turns the same missing mass into a difference of up to δ. So δ is the matching tolerance on μ_2 itself, tighter than the 0.01 the pairing below uses, and Proposition 3 is what turns a comparison at that tolerance into one at equality. Write ϵ for the certified bound on the difference in the true second-marginal error, the dropped mass plus what the solve left in the estimate, and for two drafters a and b at one position write

$$G = \max\{0,\ D^a_{\mathrm{rec,lo}} - D^b_{\mathrm{rec,hi}},\ D^b_{\mathrm{rec,lo}} - D^a_{\mathrm{rec,hi}}\}, \qquad \text{and, where } G > 2\epsilon, \qquad \frac{G - 2\epsilon}{\tau_{2,\mathrm{hi}}} \tag{8}$$

for the difference in reach that survives the equalisation. Both G and 2ϵ are in D_{rec}, and only the second expression is in reach. G is positive on 133 of the 204 pairs, at 23 positions and all 12 prefixes, at a median 0.065 against a 2ϵ of at most 0.017. It exceeds 2ϵ on 120 of them, at 22 positions and 11 prefixes, where equation 8 is 0.108 at the median, 0.001 at the smallest and 0.476 at the largest, which is what Figure 3 draws. Over all 204 pairs the difference between two midpoints has median 0.099, ninetieth percentile 0.298 and largest 0.549, and the spread across the three families at one level has median 0.163 and largest 0.549, above the preregistered threshold on 55 of 72 comparisons. Pairing at a tolerance keeps all five families: 49 pairs at 23 positions whose μ_2 differ by at most 0.01 give a median reach difference of 0.041, a ninetieth percentile of 0.181 and a largest of 0.557.

The results are organised by $e = \mathrm{reach} - f(\mu_2) \approx \alpha_{\mathcal{P}} + g_{\mathrm{train}}(H(P_1)) + \epsilon$, where $\alpha_{\mathcal{P}}$ is the offset the prefix distribution $\mathcal{P}$ carries, g_{train} is an unspecified training-dependent term in the target's first-position entropy, and ϵ the variation left over. That expression is an organising hypothesis and nothing here is fitted to it. Its terms rest on different evidence, f on the fixed-target intervention and g_{train} on a rank diagnostic fixed in advance. The three evaluation sets do not separate $\alpha_{\mathcal{P}}$ from g_{train}, because they differ in $H(P_1)$, and inside a band they share, the offset is no longer resolved.

The main text refers to these readings. On the sweep's own scale the native second-position head has mean residual $+0.0020$ with interval $[-0.0223, +0.0231]$, truncation -0.0632 with $[-0.1039, -0.0260]$ and sharpening -0.1323 with $[-0.1761, -0.0841]$. Against the curve the pre-registration locked, the seeds agree on 14 of the fifteen checkpoint verdicts rather than 15, and three intervals failing to exclude zero is not three measurements of one effect, which is why the continuous reading carries the section instead: the seed spreads against the excursions across the five checkpoints say the three runs trace one shape rather than three. On the frozen curve the internal seeds are still positive at step 42 while both arrive at step 173, so the difference in settling step belongs to the calibration and not to the runs. Appendix F reads the same trajectory on both curves, and neither the sign nor the gap to the native head comes from the respecification. On the positions the native head reads, median $H(P_1)$ is 2.97 and 3.11 on the prose sets against 1.41 on the code holdout, with τ_2 rising with it at Spearman $+0.79$ to $+0.84$; those populations are themselves selected on that covariate, at Spearman $+0.51$ to $+0.86$, so any separation between them is read after selecting on it. In the native second-position head the entropy coupling is close to the same strength on all three distributions, between 0.552 and 0.573.

Single-family predictions inside Medusa's μ_2 window span 0.245 to 0.443 against an interval 0.079 wide, and inside the multi-token-prediction model's, 0.574 to 0.745 against one 0.030 wide. Medusa's window carries 0 native-head records, so it holds only perturbations: weighting its 4 families equally gives 0.347, and the two dispersing families alone give 0.438 over 21 records, above Medusa's own interval. The multi-token-prediction window holds 11 native-head records beside its 5 perturbation families. Weighting those groups equally gives 0.683, and the native-head records alone give 0.704; the deployed family's measured interval $[0.706, 0.736]$ excludes both, below. Table 6 puts a real head with one family and one only: the native head's interval overlaps the family interpolating toward the target's own second marginal and is disjoint from the other four. That family holds 21 records inside the multi-token-prediction window and predicts 0.745 there, which that family's interval excludes, and 0 records inside Medusa's, where the composition a real head resembles cannot be read at all.

The curve's support runs from 0.025 to 0.675. Held out against the sweep on 110 of 128 positions, with the same target, conditional rows and τ_2 throughout, the second head of the shorter run reads -0.0021 with $[-0.0244, +0.0190]$ and the native head $+0.0060$ with $[-0.0131, +0.0248]$, neither excluding zero. In all 13 of the 13 μ_2 bins holding more than one family, no two families sit on the same set of positions, so a statistic compared across families inside a bin carries prefix composition as well as error geometry. Sections 5.1 and 5.2 report the general estimator against the frozen

calibration curve, on which the perturbation sweep and the held-out trained head were measured; Sections 5.3 to 5.5 report the specialised estimator against the respecified curve, and against the frozen one where a preregistered reading was locked to it. Row selection uses a budget of 2×10^{-3} on the dropped first-token mass with a cap of 16384 retained rows, except in the fixed-target sweep, which ran at 5×10^{-3} with a cap of 8192. What that budget buys is not comparable across prefix distributions. On the positions a residual is read on the median retained row count is 2577 on the internal set, 2617 on arXiv and 633 on the code holdout, and on the positions the rule keeps it is 2409 for the multi-token-prediction family against 19 for Medusa, whose first-position laws are concentrated enough that the budget is met by a handful of rows. The code holdout, which carries the one significant difference between sets, is the least well resolved of the three. Over all of a set's positions rather than the kept ones, median $H(P_1)$ is 2.62 on the internal set, 2.78 on arXiv and 0.18 on the code holdout, so the rule narrows the separation Section 5.4 reports rather than creating it. The informativeness rule is also not the same instrument on the two deployed families: Vicuna has no second-position head, so on Medusa the width is read from Medusa's own certificate rather than from a native reference, and the 21% and 70% of Section 4.1 are quoted on a population that on the Medusa side its own certificate selected. At one of the sweep's 32 positions the cap bound before the budget did, so at its 15 records the dropped mass reached 8.4×10^{-3}. Every certificate is computed from the mass its own row set actually dropped rather than from the budget, so the wider selection widens those intervals instead of invalidating them. The informativeness threshold is 0.10, applied to the width of the certified interval for D_{rec} divided by the midpoint of the certified interval for τ_2. A relative width on its own would admit a position whose width is zero because the estimator clipped both ends of the interval into $[0, 1]$, where a single retained row leaves nothing to identify no matter how narrow the interval looks. The rule also requires more than one retained row, which is a statement about what a position can answer rather than a threshold: 0 records now pass it that way, against one position and sixteen records before the requirement was added. Dropping that position moves the code holdout from 98 kept positions to 97 and the two step-3000 arms that read it by about a thousandth, and leaves every verdict, every deployed-family population and the sweep's statistics unchanged. Before any checkpoint is read, the internal set's native head has to reproduce the committed measurement of the same positions. It does, to $+4.77 \times 10^{-5}$ in μ_2 and -1.44×10^{-4} in reach against a tolerance of 0.01. That reproduction exists for the internal set only, whose positions were measured in an earlier committed round; the code holdout and arXiv have no prior measurement of the same positions, so on those two the harness checks that this gate passed and reads the trajectory without a reproduction of its own. That gate compares means, and its reference is the cross-head round, which holds the same frozen sequences the internal set does, so neither a different statistic nor a tighter tolerance in it could have found the deployed row's divergence. What locates that is the retained row count read across the two rounds, and what compares record against record is the SHA-256 manifest. A prompt on the internal set and an item on the external sets are the unit of replication, so every interval on a mean residual is a cluster bootstrap of 4000 draws over those units rather than over positions, of which the sweep has 13 and the cross-head 16. A draw takes the clusters with replacement to their original number and averages over every position in the clusters it drew, and the interval is the central 95% of those draws, cut at the 2.5 and 97.5 percentiles, so each interval reported is one reproducible realisation of the bootstrap distribution rather than an average over its randomness. The sets reach similar values at different marginal errors. On the positions each native head is read on, mean μ_2 is 0.172, 0.208 and 0.148 against reach 0.721, 0.667 and 0.689, so in reach the code holdout is not the lowest of the three; at step 3000 its first seed sits at $\mu_2 = 0.293$ against 0.382 and 0.398 on the prose sets, and reading a similar reach at a lower marginal error against a falling curve is what makes its residual the most negative. Between the two calibrations 4 of the forty-five run and checkpoint verdicts change, the code seeds splitting at step 42. One of the verdicts Section 5.3 counts sits inside that bootstrap's own resolution, as does one reading on the preregistered curve: re-drawn under 24 seeds, the internal set's second seed at step 42 turns over in 6 of them on the respecified curve and the code set's second seed at the same checkpoint in 3 on the frozen one, each of which would move a reported count by one. A difference between two prefix distributions cannot be paired, because the two sets hold different positions; those intervals come from an unpaired cluster bootstrap that draws each set's clusters independently and takes the difference of the two means on every draw. The differences between drafters at one position are paired, and are marked as such where they appear. The rank correlations between $H(P_1)$ and the residual are read from midpoint residuals like every other residual here, and the certificate does not pin their level. The rise does not depend on where inside those intervals the two checkpoints are read. Taking both at the same relative position inside their own intervals, over 297 readings on a

grid that runs reach and μ_2 each from one endpoint to the other across the nine run and set pairs, the smallest rise from step 10 to step 721 is $+0.317$. That sweep moves every position the same way, so it rules out a common-mode reading of the rise and not an adversarial one: perturbing positions independently inside their own boxes can weaken, strengthen or reverse a rank relation, and no certificate bounds what it would do here.

The re-expression of the calibration rebuilds the 14 knots from the same frozen records, the same native-derived mask and the same bins under the specialised estimator. The bins are the fixed intervals of width 0.05 in μ_2, with a knot at each center, so the same bins means the same edges and not the same members: a position whose μ_2 moves under the re-expression moves with it. Each knot is the unweighted mean reach of whatever drafters fall in its bin, and which families those are changes from bin to bin, so the curve averages over the very spread Section 5.1 measures. The knots at the top of the support rest on the fewest rows. Run with the general estimator the procedure returns the original knots to 0.00043, inside the 0.00050 rounding with which they were written, and only then is the estimator changed. All 14 shifts are positive, between $+0.0039$ and $+0.0123$ as rebuilt, and the knots are then written to three decimals for use, which is the rounding the reverse control above is read against.

F THE SUPPORT EDGE OF THE CALIBRATION CURVE

This probe is post hoc. It was not preregistered, none of its numbers appears in a result table, and it exists because the code holdout's residual turned out to depend on where in the support its positions sit, which would otherwise have been read as a statement about the domain. It is run on both calibration curves, because a check read only on the frozen curve cannot rule anything out about the respecified-scale residuals of Section 5.4.

The curve is a linear interpolation through 14 knots. Its first segment falls steeply, so near the bottom of the support a small error in μ_2 produces a large error in f, and a set whose positions pile up there inherits a residual from the geometry of the curve rather than from the head. The code holdout piles up there: the target is a code model reading code and is nearly right at the second position at many of its positions.

The test restricts the residual to the interior of the support and asks which residuals survive. On the code set, over the full support the native second-position head has mean residual -0.0570 with interval $[-0.0955, -0.0231]$ over 84 positions, which excludes zero; restricted to $\mu_2 \in [0.10, 0.50]$ it is -0.0152 with $[-0.0493, +0.0161]$ over 58 positions, which does not. The first seed's trained head at step 3000 behaves the other way: -0.1021 with $[-0.1467, -0.0616]$ over 88 positions on the full support, and -0.0938 with $[-0.1403, -0.0502]$ over 73 positions in the interior, both excluding zero.

The same restriction on the respecified curve, which is the scale Section 5.4 reports, gives the same answer, and reads all 3 seeds rather than one. The native second-position head has -0.0678 with $[-0.1064, -0.0339]$ over 84 positions on the full support and -0.0257 with $[-0.0600, +0.0056]$ over 58 positions in the interior, which no longer excludes zero. Every seed at step 3000 keeps an interior residual that does exclude zero, between -0.1029 and -0.0727.

So the native second-position head's offset on the code set does not survive the restriction and the offset the trained heads acquire does. The restriction changes which positions are read as well as the window, so it separates the two heads without deciding whether the native offset comes from the curve's geometry or from the positions that carry it. Only the offset the trained heads acquire is read as a result in Section 5.4.

The same support rule also moves the sample along training, because positions leave the support when μ_2 falls outside it and training pulls μ_2 inward. The first seed's internal set drops 15 positions at step 10 and 2 at step 3000, the code holdout 26 and 10, and arXiv 17 and 7. Restricting to the positions in support at every checkpoint and every seed holds the sample fixed, at 93, 60 and 149 positions. On that subset the code holdout is significantly negative from step 42 onward. A second and different restriction, to the window between 0.10 and 0.50, makes a prose set distinguishable from the curve, which is the reverse of what the full support shows and of what an earlier version of this appendix reported: in that window the trained heads of the arXiv set at step 3000 exclude zero on the negative side at every seed. The window is the reason. It removes the top of the support,

and on the prose sets that is where their high-entropy positions sit, so what the restriction leaves is their low-entropy tail. Splitting inside this one window by $H(P_1)$ moves the arms apart without separating them: of the 6 prose set and seed arms whose positions sit above 1.5 in $H(P_1)$, all 6 include zero, while 3 of the 6 arms made of the positions below it exclude zero on the negative side, so half of the low-entropy arms do and half do not. Holding the window fixed bounds how much of the comparison the curve's shape can carry, but not all of it: two sets can occupy one window and still sit at different μ_2 inside it, and f is not linear there. The probe computes one window narrower still, between 0.10 and 0.30, and there all 9 of the 9 set and seed arms exclude zero on the negative side, the code holdout no longer lowest among them. No ordering can be read from that window: it keeps as few as 13 positions per arm, and what it keeps is each set's own low-entropy tail, which is the same selection the previous paragraph describes. On the common-position subset above, 14 of 15 checkpoints keep all three seeds on one verdict rather than 15, the exception being the internal set at step 3000. Post hoc.

Restricting the code set to $H(P_1)$ between 1.5 and 4.5 moves its three step-3000 means from -0.1100, -0.0896 and -0.0894 to -0.0298, -0.0126 and -0.0153, putting all 3 across zero on 32, 31 and 31 of the set's positions, at widths no more than 1.05 times their unrestricted ones. The point estimate moves by a factor of four to seven and the interval width does not, so this is not a loss of power. That is one window: moved down to 1.2 the three seeds exclude zero again. The lower edge was swept, and that flip is a single point rather than a trend, so the restriction locates the offset in the low-entropy positions rather than establishing its absence. At step 173 the nine entropy correlations run from $+0.125$ to $+0.450$. On the respecified curve the internal set stops being distinguishable by step 42 and arXiv by step 173, and neither moves away through step 3000. Against the frozen curve the code set's three runs at step 3000 read -0.1009, -0.0809 and -0.0807 against a native head of -0.0570. The support restriction also removes positions before any trajectory is read, and the three sets do not lose the same amount. Of the positions the informativeness rule keeps, 0 are dropped at the native head on the internal set, 3 on arXiv and 13 on the code holdout, leaving the 110, 170 and 84 that head is read on. At a trained head the drop is larger and moves with the checkpoint, because training pulls μ_2 across the support edge: across the five checkpoints and three seeds a trained head is read on 94 to 108 internal positions, 154 to 167 on arXiv and 69 to 90 on the code holdout. The code holdout loses the largest share and loses it at the bottom edge, which is where the first paragraph of this appendix says the curve's geometry manufactures a residual.

The curve is the per-bin mean of whatever drafters land in the bin, and the mix is not constant along μ_2: 1 of the 14 bins holds a single family, and the families that concentrate mass are the only ones present at the top of the support. Weighting the families in each bin equally instead moves the knots by between -0.033 and $+0.052$. Carried through the same respecification, the native second-position head on the internal set reads $+0.0199$ with $[+0.0009, +0.0385]$, which excludes zero where against the curve as fitted it straddles it. That lower end is within a thousandth of zero, and which side of zero it falls on is set by the draw as much as by the weighting: over 41 draws of the same bootstrap it runs from -0.0010 to $+0.0014$ and 14 of them put it at or below zero. arXiv moves the same way and does not cross: $+0.0122$ with $[-0.0041, +0.0276]$. The code holdout stays negative at -0.0500 with $[-0.0897, -0.0156]$. So the sentence in Section 5.4 that the native head is indistinguishable from zero on both prose sets holds for the curve as fitted and not for that weighting, while each set's native reading moves by about $+0.018$, so no gap between two sets moves by more than 0.0006 and the ordering is unchanged. Between 0.525 and 0.575 the equal-weighted knots rise, where both fitted curves fall throughout, because the higher of those two is held up by 11 records. No position any native head is read on lies there, so the reading above is unaffected. The preregistration locked the curve as fitted, and this weighting is post hoc.

The frozen result document shipped with the supplement gives the family table of Table 6 on a computation that clamped a query outside the knot range to the nearest knot instead of refusing it, so three of its rows differ from the table here: the family running toward the second marginal reads 72 records and $+0.0068$ here against 100 and $+0.0116$ under the clamp, 28 of them read off the lowest knot, and truncation and sharpening move in the fourth decimal. That family runs to $q_2 = P_2$, where μ_2 goes to zero, so its records are the ones that leave the bottom of the support. The document also asserts that no value in it is read off a clamped end, which is true of its first reading and false of that table. The direction of every family is unchanged, the frozen curve is fitted on per-bin means of reach rather than on residuals and reproduces from the records either way, and no verdict in this

paper turns on it. A correction is appended to that document in the supplement rather than folded into it.