SAME CODE, DIFFERENT REACH: LEARNING POSITIONAL COMPARISONS FOR LENGTH EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Length extrapolation in copying requires reliable source retrieval beyond the training range. We study how positional structure and training shape this learned computation through controlled experiments in small transformers. Across six harmonic frequency grids and twelve paired seeds, 25% frequency-plane dropout improves correct-history token accuracy at four times the training length by 34–41 percentage points on four grids; one shows no resolved gain and one worsens. All bands are restored at evaluation, so these gains arise with the positional code held fixed. To investigate the learned addressing computation, we distinguish context matching from selection among matching occurrences. Positional features help construct retrieval contexts as well as distinguish addresses. A task-derived contrast predicts mean spectral-growth peaks at four training lengths, but different learned band mixtures implement similar local comparisons with divergent distant responses. Removing distant attention competitors improves context selection and copying. Frozen short-range spectral forecasts locate consequential distant attention, although equal-mass controls show no resolved accuracy advantage for the exact selected addresses. Matched continuations and signed-coefficient transfers establish a contributing role for learned positional comparisons. Band interventions reveal useful alternatives, and a conditional subset-margin theorem gives sufficient conditions for distant separation in a reduced model after all bands are restored. Context discrimination improves, while ambiguity among matching occurrences remains. Together, these findings distinguish a positional code's available comparisons from those learned through training, and identify support-dependent positional regularization as a promising approach to length extrapolation.

1 INTRODUCTION

Harmonic positional encodings supply recurring features. Learned attention scores turn those features into preferences between token positions. Yet models given the same frequencies can copy short sequences equally well and diverge sharply on longer inputs. We ask: *which learned positional comparisons remain useful beyond the training length?* Longer inputs expose these comparisons to distant recurrences and additional attention competitors.

Copying makes this question concrete because each output has a known source address. Context-based retrieval matches the preceding output tokens to source contexts (Olsson et al., 2022; Jelassi et al., 2024). When several locations share that context, the model must also identify the intended occurrence (Wen et al., 2026). We distinguish these two demands through independent context edits and positional shifts. Positional features affect both: they help gather nearby tokens into a context representation and distinguish otherwise similar addresses. A model may therefore benefit from positional training because it preserves context matching, even while repeated occurrences remain ambiguous.

Our first finding connects task geometry to spectral learning. A contrast between the local predecessor and the repeated token in the copy predicts group-level spectral-growth peaks in learned absolute positional embeddings (APE). But this preference does not determine a unique spectrum. Bands compensate to produce similar local comparisons, and fixed-width frequency grids that agree

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.

locally give different distant responses. Removing distant first-layer competitors improves context selection and copying. Frozen short-range fits select consequential distant attention; equal-mass controls connect the rescue to attention allocation during context construction.

Our second finding is that training can change this behavior with the code held fixed. At a 25% rate, dropout of complete harmonic bands improves correct-history token accuracy at four times the training length by 34–41 percentage points on four of six grids, with negligible short-accuracy changes on those four. The other grids show no resolved gain or an adverse effect. Matched continuations and signed-coefficient transfers test the contribution of learned positional comparisons; controlled probes show stronger context discrimination without reliable occurrence selection.

Finally, we test why training under band removal can help when all bands return at evaluation. Conditional band interventions reveal useful alternatives, and a subset-margin theorem shows when subset competence and complementary distant separation protect the restored code in a reduced model. Together, these results link task demands, learned harmonic comparisons and length extrapolation.

2 Task and measurements

A causal transformer receives a source string of length L over a vocabulary of V tokens and reproduces it after a separator. Unless specified otherwise, training uses $L \leq 120$, evaluation extends to $L = 480$, and the fixed-code models have width 64, two heads and two blocks. The main fixed-code experiments use additive harmonic features $\{\cos(\omega_b i), \sin(\omega_b i)\}_{b=1}^{B}$ at position i; each sine/cosine pair is a band. Band-count comparisons preserve model width. Learned APE uses cyclic spans; the main fixed-code cohorts use nonwrapping spans with random origins.

We distinguish embedding Fourier power, the signed first-layer positional score $K(d) = \sum_b [c_b \cos(\omega_b d) + s_b \sin(\omega_b d)]$, and native attention-score margins after recomputing the complete network. The stationary diagnostic K omits content terms and input-dependent normalization. A band's native utility is the signed margin decrease when it is removed; this need not equal its power or coefficient amplitude.

Component dropout masks complete positional bands during training, restores the retained code's RMS, and reinstates every band at evaluation. We compare ordinary and dropout training to expose which comparisons change, then test their contribution with targeted interventions. The primary behavioral measure is content-token accuracy with correct preceding outputs. Attention selection and margins are separate diagnostics. Main intervention cohorts use twelve paired seeds; uncertainty is over trained seeds, averaging heads and probe inputs within each model. Acquisition, full-budget and continuation cohorts remain distinct (Appendices A and G).

3 Copying requires context matching and occurrence selection

To understand which learned comparisons must survive extrapolation, we first separate the two demands of copying. An early layer can gather preceding tokens into context features; a later comparison matches the output context to a source candidate and retrieves its continuation (Elhage et al., 2021; Olsson et al., 2022). When candidates share that context, matching alone cannot identify the intended occurrence. We separate the effects of context and position by controlling them on different axes.

Change context and positional gap independently. Our crossed probe uses correctly copied L240 strings. The correct source and a wrong-payload rival, 96 tokens apart, can share up to 64 preceding tokens. At fixed positional coordinates, changing the rival's matching prefix measures the effect of context matching. At fixed tokens, shifting the separator and copied segment by a PE gap $g = 0, 16, \ldots, 512$ measures the effect of position. Source length, candidate addresses, all token history and within-segment spacing remain fixed across gaps. Changes across this sweep therefore identify the response to position with the available text held constant, including when the candidate contexts match.

Measure the two addressing demands on the same gap sweep. Let G contain the correct source and the planted rival sharing its context. Attention conditional on the original source positions

satisfies

$$P(\text{correct} \mid \text{source}) = P(G \mid \text{source})P(\text{correct} \mid G). \tag{1}$$

With correct-minus-strongest-rival margins outside and inside G, $M_{\text{total}} = \min(M_{\text{out}}, M_{\text{in}})$. Both identities hold per example, head and gap. Figure 1 compares context-class selection and discrimination within the matching class under the same positional perturbation. Dropout raises gap-averaged class probability from **40.1% to 68.3%**, while correct attention within the class is **56.1% and 48.0%**. The fraction limited by an outside-class rival falls from 64.8% to 30.4%. Positional training can improve context selection without reliably resolving the matching-context occurrences.

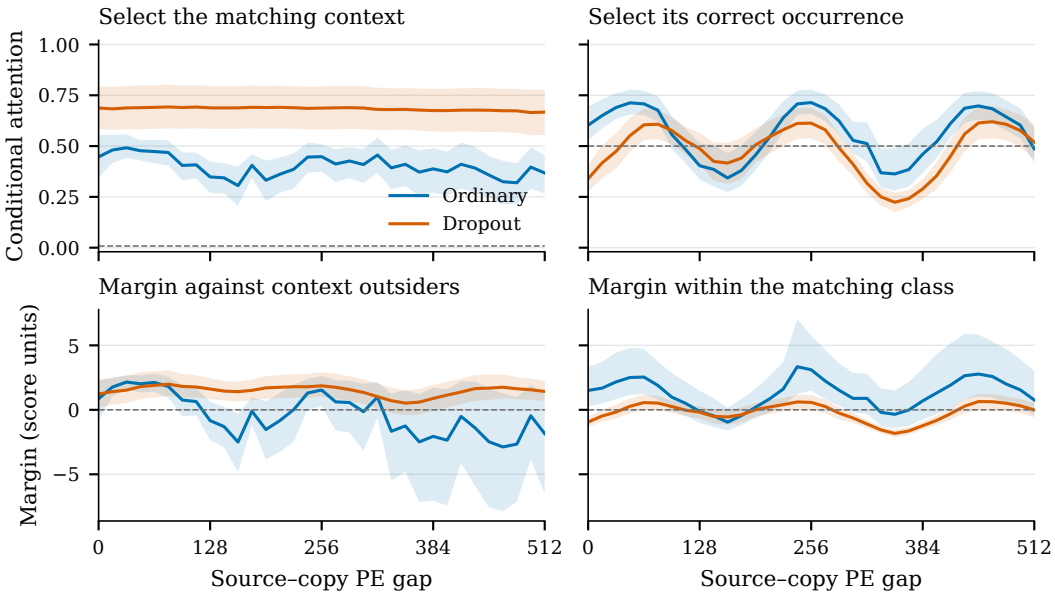


Figure 1: **Better context selection does not ensure correct occurrence selection.** Top: attention to the matching-context class and to the correct member within it. Bottom: correct-source margins against the strongest outsider and the matching-context rival. The same L240 strings, 64-token matched prefixes and candidate pair are used at every gap; only copied-side PE coordinates shift. Dashed lines mark uniform attention (top) and zero margin (bottom). Twelve seeds per recipe, 32 backgrounds and both heads averaged within seed; pointwise 95% seed-bootstrap intervals.

Position changes the preference between matching occurrences, but the preference need not favor the correct one: correct ranking within the pair averages 51.9% ordinary and 48.8% dropout. Removing the $\omega = 0.03162$ band lowers ordinary ranking to 31.3%. A frequency can contribute useful positional discrimination without making the complete occurrence choice reliable.

Context matching remains useful against distant rivals. A complementary probe varies physical rival displacement in both directions. At each displacement, let M_m be the correct-minus-rival score with m matching preceding tokens. Then $C = M_0 - M_{16}$ measures the context-mismatch advantage at that same displacement. At L120, ordinary models have a mean context advantage of **62.2 score units**, compared with a **2.26-unit** matched-context preference. At L480, ordinary context advantage falls to **36.9**, while dropout preserves **62.6** (Appendix D, Figure 9). The context window and positional gap are different scales: a local context cue can distinguish candidates far apart.

Positional features support that context cue as well as occurrence selection: joint band removal reduces context-response RMS from 32.43 to 1.06 in ordinary models and from 28.86 to 0.35 with dropout. The controlled effects therefore need not correspond to separate neural modules. Repeated-prefix tests also separate token success from address selection: exact period-16 repetition gives 99.8% accuracy at L120, but changing the payload after each shared 15-token prefix lowers it to 56.8%/55.7%. At L480, dropout can exploit information beyond a shared prefix, although ambiguity remains in other configurations (Appendix E). These results locate a central extrapolation demand: preserve the context comparison even as more source candidates become available. The next section asks which harmonic features training recruits to construct that comparison.

4 COPYING DEMANDS SHAPE SPECTRAL ACQUISITION

Context-based copying gives harmonic learning a concrete target: gather local information for a later source comparison. We ask whether this geometry predicts which bands learned APE acquires. The prediction concerns an early spectral preference; it does not specify the complete copying algorithm or resolve every repeated occurrence.

Length sets a task-relevant contrast. Predecessor context supplies an induction cue. Copying places the corresponding repeated token at lag $L + 1$, while the local predecessor is at lag 1. On a 512-position cycle, a cosine band f centered at the predecessor separates these relations by $1 - \cos(2\pi f L/512)$. Averaging over uniformly sampled source lengths, weighted by their number of copied tokens, gives

$$D_f(T) = 1 - \frac{\sum_{L=1}^{T} L \cos(2\pi f L/512)}{\sum_{L=1}^{T} L}. \tag{2}$$

Maximizing this contrast over bands 1–64 predicts peaks at **6, 4, 3 and 2** for maximum training lengths $T = 60, 90, 120, 180$. All four measured mean growth peaks agree (Figure 2), without fitted coefficients. The prediction was fixed before these runs; power growth is the secondary readout, following early score demand. Agreement extends beyond the peak: across all 64 bands, Pearson correlations with seed-mean profiles are **0.52–0.80 for early score demand** and **0.72–0.82 for subsequent power growth**. It also captures 60–87% of adjacent-band rise/fall directions in the early-demand profiles. These results support a task-dependent spectral pattern, with less reliable exact maxima: growth peaks match in 11/16 seeds and early-demand peaks in 7/16. Appendix B reports the full-profile metrics and baseline comparisons.

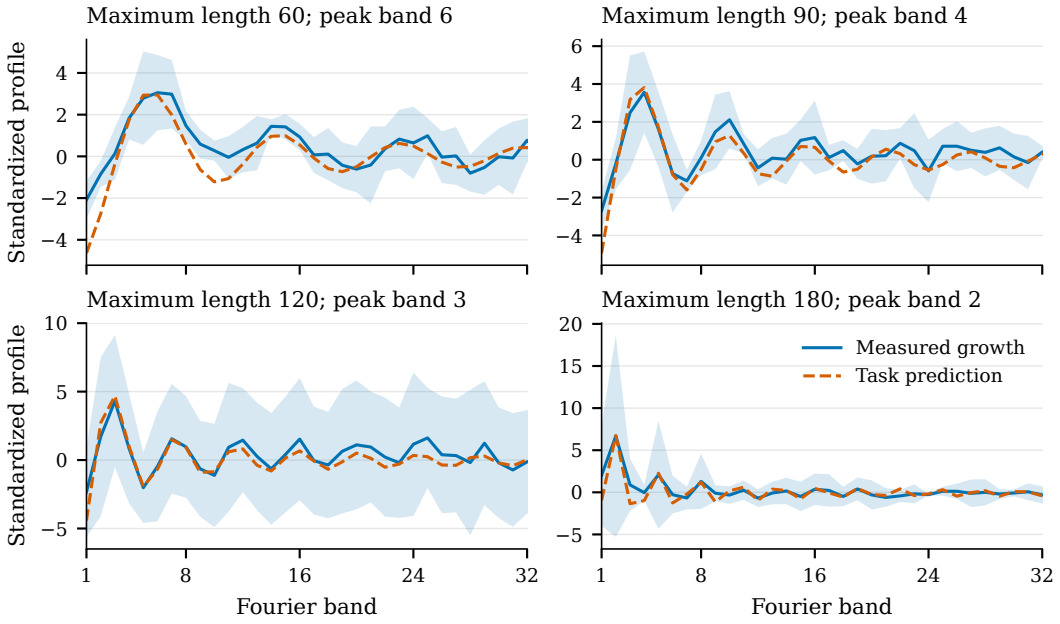


Figure 2: **Longer copying tasks favor longer spectral scales.** Measured early log-power growth and the task-only prediction at four maximum source lengths. Four seeds per length; 95% seed-level intervals. Profiles are standardized; peaks are selected over bins 1–64, with bins 1–32 displayed. Vertical scales vary. Growth is measured between updates 512 and 1,536.

The comparison is more stable than its spectral realization. Across 36 learned-APE models, final signed Fourier-coefficient variation is **67–105%**, whereas variation in short-range positional comparisons is **11–22%** (relative RMS). Retaining between-band covariance reduces comparison variance by 73–86%. Different bands compensate to preserve a similar comparison. Controlled edits support a functional version of this observation: coordinated amplitude changes give held-out token KL of **0.0329**, versus **0.1148** for randomized-sign controls, with accuracy losses of 0.98 versus

3.05 points across eight sinusoidal models. Appendix C gives the interventions and retrospective uncertainty analysis.

Task geometry thus constrains what the learned comparison must do without specifying a unique band mixture. Short-range agreement leaves its distant behavior open, which we test next with fixed frequency grids.

5 Local agreement does not prevent distant interference

At fixed width, we compare the local neighborhoods used to construct context with the distant competitors against which that context must remain discriminative.

Shared local comparisons, different distant extensions. Ordinary 4/8/16-band models use the same frequency endpoints, redistributing nodes as the band count changes. Their first-layer positional score shapes agree closely over lags 0–16: paired centered cosines are **0.971 for 4/8 bands and 0.987 for 8/16**. Over lags 121–480, agreements fall to **0.132 and** -0.217 (Figure 3). The native context responses also agree locally: ordinary 8/16-band models reach 90% of their measured prefix response within **6.60/7.21 tokens**, with normalized curve RMS difference 0.049. Local functional agreement thus coexists with substantially different distant positional scores.

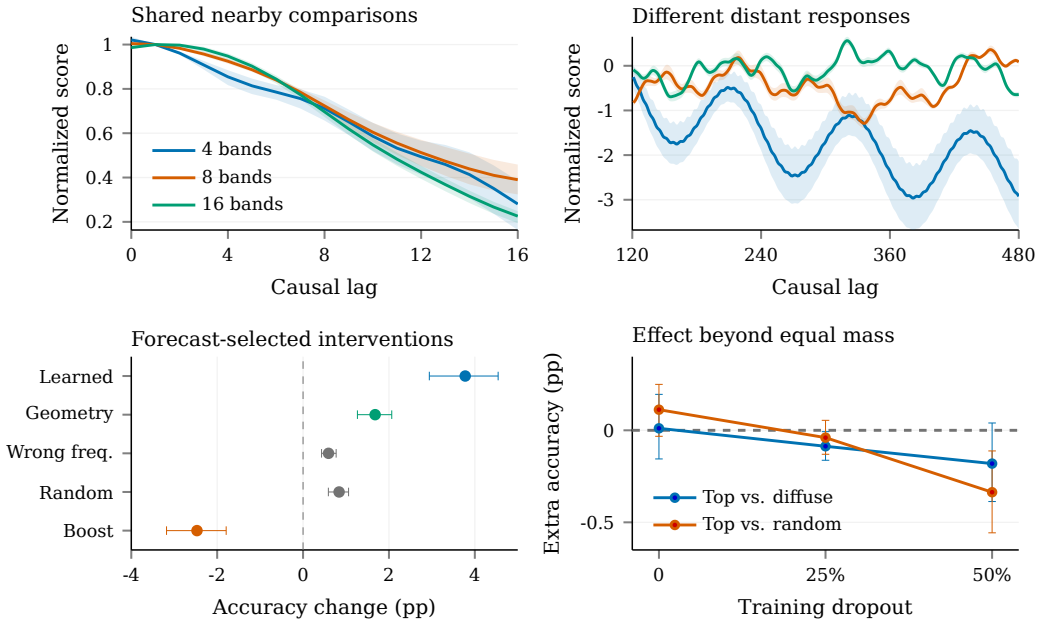


Figure 3: **Learned comparisons locate consequential distant attention.** Top: stationary positional scores normalized by each seed's $K(1)$; width, frequency endpoints and $V = 4$ are fixed. Bottom left: actual output effects of 12-key interventions selected by frozen forecasts, ordinary models across six grids. Bottom right: learned-key removal minus equal-mass diffuse or random removal. Fresh L240 strings, three output offsets, four origins; twelve paired seeds, averaging grids within seed. Intervals are 95% seed bootstraps. Equal-mass controls use native attention for the common intervention dose.

Distant interference disrupts context matching. Removing only distant first-layer keys raises ordinary L480 token accuracy from **65.0% to 93.7%**, versus 78.5% under count-matched random removal; short outputs are exactly unchanged. A fresh probe also raises matching-context selection from **32.3% to 46.0%**. These interventions connect a distant response to the context-selection failure identified in Section 3. More specific controls distinguish harmful distant content from useful boundary markers; four-band models can rely on the latter (Appendix L). Long-range attention can therefore help or harm, depending on what it carries.

Short-range forecasts select behaviorally relevant competitors. Across 216 existing models, query-conditioned signed harmonic fits use only training-supported lags 1–240 to forecast first-layer scores at lags 243–480. They use the known frequencies and short score observations, not long scores or copying outcomes. At fixed absolute coordinates, these are new-lag forecasts, not tests on unseen positional coordinates. Frozen forecasts select 12 distant content keys per head for removal at one output query, preserving marker/self attention and redistributing the removed mass to near content. On fresh L240 inputs at three output offsets, ordinary-model accuracy improves **3.78 points** [2.94, 4.54], versus 1.68 for geometry-only selection and 0.84 for random keys; strengthening the learned selection costs 2.47 points. Removing the *same mass* proportionally across all distant content gives 3.77 points: the paired difference is 0.01 [−0.16, 0.20]. The forecast thus locates consequential distant mass; exact selected addresses have no resolved accuracy advantage at equal mass. Rescue shrinks after dropout. All grids, failed forecasts, dose controls and the original single-offset test are retained in Appendix N. We next test how training changes these comparisons with the code fixed.

6 The same code can support different extrapolation

Compensation shows that several band mixtures can support short copying; distant interference shows that their extrapolation need not agree. Component dropout trains with whole bands removed while retaining the copying objective. All bands return at test time, so paired models receive the same positional code. We first measure changes in behavior and context discrimination, then intervene on the learned comparisons to test their contribution.

Behavior and context advantage across complete cohorts. A study crosses six grids, rates $p = 0, .25, .5$, and twelve paired seeds: 216 models at $V = 4$. At $L = 480$, 25% dropout improves correct-history content accuracy by **34–41 percentage points** on Standard8, Standard16, Base256-16 and Geometric16. Their short-accuracy changes are below 0.1 points in magnitude. All four grids also improve at 50%, with all eight positive intervals excluding zero. Standard4 has no resolved gain; Geometric8 worsens at both doses (Figure 4).

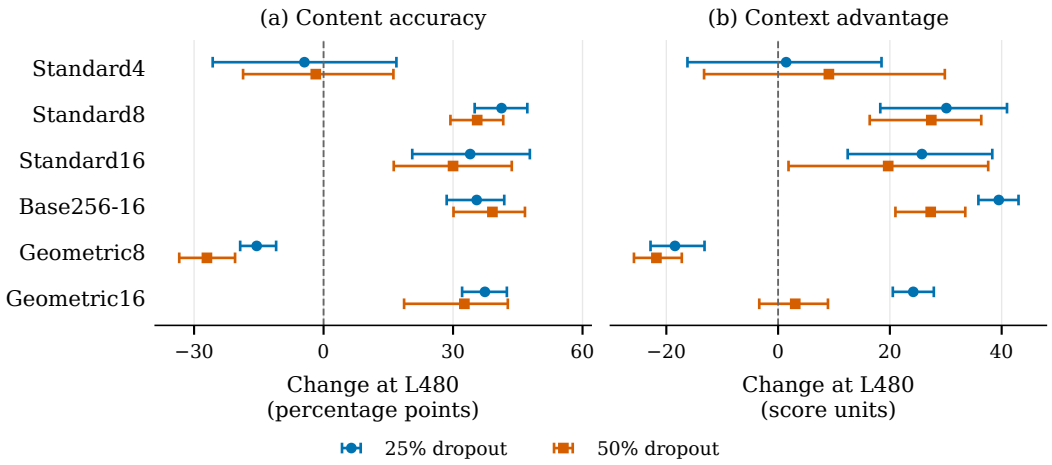


Figure 4: **Fixed-code gains largely track preserved context discrimination.** Dropout-minus-ordinary L480 changes in content accuracy (left) and context advantage $C = M_0 - M_{16}$ (right). All six grids and both doses are retained. Context advantage averages balanced rival displacements, backgrounds, origins and heads within each seed. Paired 95% seed-bootstrap intervals over twelve seeds, without multiplicity adjustment.

At 25%, the four favorable bases gain **24–39 context-score units**, whereas Geometric8 loses **18.45**. At 50%, Geometric16 improves copying without a resolved context-advantage increase. This diagnostic explains a common pattern, not every gain. Additional vocabulary and custom-code cohorts, confirmation seeds, mixed V2 outcomes and limited RoPE gains are retained in Appendix G.

6.1 Repair and transfer test the contributing comparison

Matched continuations separate repair from initial acquisition. From the same ordinary parents, equal 8,192-update continuations give Standard16 L480 accuracy of **71.18% ordinary versus 97.73% dropout**; both arms average **99.82%** at L120. The paired gain is 26.55 points (95% CI [15.35, 38.02]). This is a separate cohort from the full-budget comparison.

Signed positional comparisons transfer part of the gain. We transfer only the signed sine/cosine coefficients of the first-layer *displacement-dependent positional QK numerator*, leaving the recipient fixed with native normalization active. Transfer size is selected and independently validated on short examples. Moving an ordinary recipient toward dropout improves L480 accuracy by **3.13 points** [1.50, 4.92]; moving away changes it by -8.71 **points** $[-11.99, -5.94]$. Position-sum coefficient transfers have approximately zero mean effect (Figure 5). These signed comparisons contribute to the behavioral gain, but their effect is partial and does not isolate occurrence discrimination from context construction.

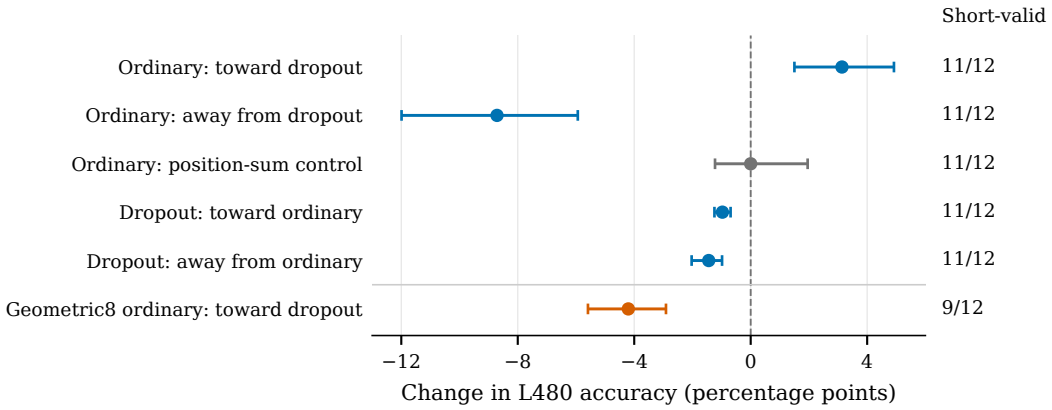


Figure 5: **Signed positional comparisons carry a transferable behavioral effect.** Mean accuracy changes and 95% seed-bootstrap intervals, all twelve outcomes per row. Counts give nonzero transfers passing independent short validation: accuracy within 0.5 points and cross-entropy within 0.02 nats of the recipient. Transfer sizes differ across directions and controls.

Geometric8 provides a directional counterexample: dropout lowers long accuracy from 38.32% to 26.63%, and transfer toward dropout lowers ordinary-recipient accuracy by 4.20 points [2.91, 5.59]. Both directions can hurt a dropout recipient, consistent with co-adapted comparisons (Appendix H).

Context construction is an intervention target. Short-only fitting that preserves clean first-layer attention under band removal improves eight-band L480 accuracy by 2.32–3.11 points with nearly unchanged short copying. In a separate 512-update Standard16 continuation, clean/masked training gives **64.5%/86.1%** long accuracy. Transferring only the learned near–far content competition reaches **82.9%**, keeping recipient values, marker weights and within-group attention fixed. These interventions connect the distant-interference diagnosis to repair; they do not establish complete mediation. Limited band budgets can incur local costs: dropout changes short context-class selection by -14.4 points with four bands and $+17.1$ with sixteen (Appendix L).

6.2 Useful alternatives help retain context margins

Training under band removal could help the restored code if several surviving combinations learn useful context comparisons and their distant failures do not coincide. We test two observable consequences of this proposed mechanism: usefulness under partial support and retention of the context advantage $C = M_0 - M_{16}$. The following theorem states the additional conditions needed for a margin guarantee in a reduced model.

For Standard16, dropout reduces single-band deletion effects while increasing conditional usefulness when other bands are absent (Figure 6). The effective spread of positive conditional utilities

rises from **6.35 to 10.24**; Base256-16 rises from **5.19 to 11.16**. Geometric8 retains similar spread, 5.28 to 5.44, while its utilities weaken. Breadth alone does not imply useful alternatives.

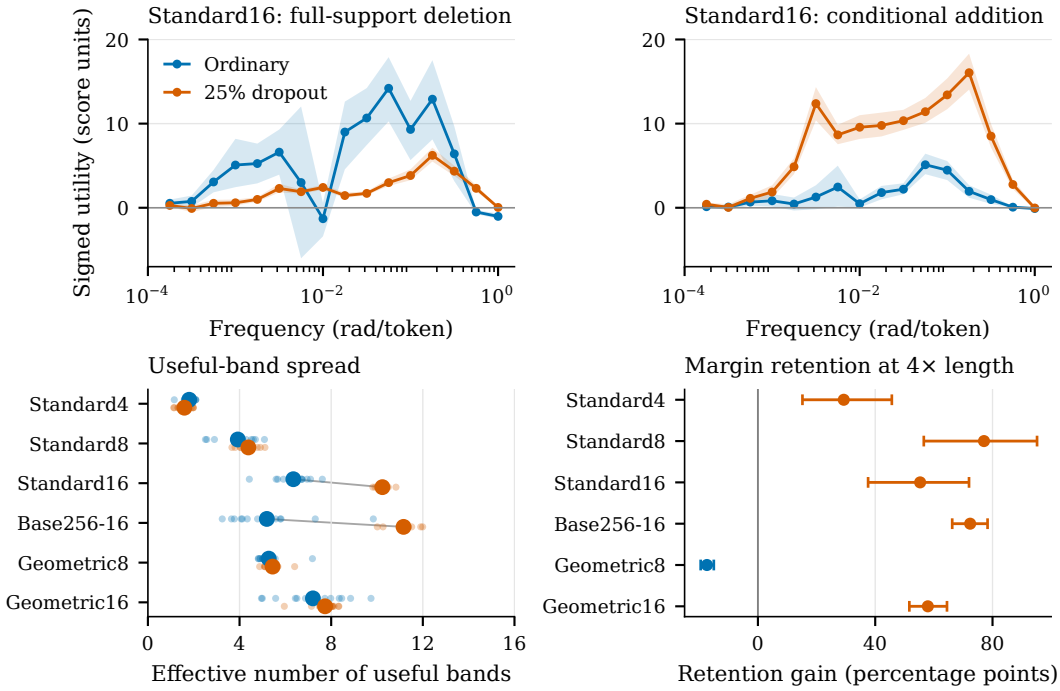


Figure 6: **Useful band alternatives accompany retained long-range context margins.** Top: Standard16 signed deletion and conditional-addition utilities; addition averages four shared random half-support masks. Bottom left: positive-utility spread $(\sum_b [U_b]_+)^2 / \sum_b [U_b]_+^2$, individual seeds and means. Bottom right: change in margin retention, using the seed-mean minimum over eight long displacements divided by the cohort short-range mean. It can exceed 100%. Pointwise or paired 95% seed-bootstrap intervals; utility and retention use separate probe panels.

Standard16's short-range context advantage is 62.20 ordinary and 59.34 with dropout. The sampled long-range minimum is 30.55 versus 61.97: retention of **49.1% versus 104.4%**, a paired gain of 55.3 points [37.6, 72.0]. These are sampled mean-margin curves, not worst-case guarantees. Standard4 also improves the ratio without a resolved copying gain; Geometric8 loses margin. Thus retained context discrimination is one requirement for copying, consistent with the separate occurrence demand established earlier.

6.3 When subset competence protects the restored code

A reduced harmonic model makes the proposed explanation conditional and precise. Let $M_S(d) = \sum_{i \in S} a_i[1 - \cos(\omega_i d)]$, $a_i \geq 0$, model aligned separation from a preferred relation. For a uniformly random k-subset of n bands, without rescaling and independent of short comparisons t, suppose

$$\mathbb{E}_{S,t}[(\gamma - M_S(t))_+^2] \leq \varepsilon^2. \tag{3}$$

If each d in an extrapolation set E has at least $g \geq k$ bands satisfying $1 - \cos(\omega_i d) \geq \kappa > 0$, then

$$\min_{d \in E} M_{\text{full}}(d) \geq \frac{\kappa g}{2k} \left(\gamma - \frac{\varepsilon}{\sqrt{q}} \right)_+, \qquad q = \frac{\binom{g}{k}}{\binom{n}{k}}. \tag{4}$$

The fraction q of masks retains only bands separating a given distant rival. Small masked shortfall prevents all such subsets from having negligible weight; restoring the full code then provides a floor. Full-support success alone can instead concentrate weight on bands that recur together at a distant address. The result requires both learned subset competence and complementary separation; it does not say that masking automatically achieves either.

Native utilities are signed and nonadditive, so they are not the theorem's coefficients. Appendix J supplies the proof, scaling convention and signed extension. A separate audit obtains positive panel-margin certificates for 28/72 dropout models and 0/72 ordinary models, using long-context residual calibration. Phase-free controls also certify margins, precluding a unique harmonic explanation (Appendix K.1).

7 Related work

Harmonic representations and learned frequency use. Sinusoidal encodings and RoPE supply harmonic positional structure (Vaswani et al., 2017; Su et al., 2024); learned Fourier maps and spectral relative-position mechanisms adapt it (Li et al., 2021; Choromanski et al., 2024). Fourier-feature and spectral-bias studies analyze representation and learning (Rahimi & Recht, 2007; Tancik et al., 2020; Rahaman et al., 2019), and modular-addition transformers learn task-structured Fourier circuits (Nanda et al., 2023). RoPE analyses connect frequency use to training range, band specialization and data distances (Barbero et al., 2025; Oka et al., 2026b;a; Wu et al., 2026). We connect copying demands to acquired additive-code spectra and to the comparisons shared by different mixtures.

Copying mechanisms and length generalization. Induction and n-gram copying explain context-based retrieval (Elhage et al., 2021; Olsson et al., 2022; Jelassi et al., 2024); generalized and soft induction analyze richer context features (Chen et al., 2024; D'Angelo et al., 2026). Repeated-context ambiguity motivates separating retrieval from occurrence selection (Wen et al., 2026). Our mechanism section uses this established distinction to interpret positional interventions. Studies of positional schemes and algorithmic generalization provide broader perspectives (Kazemnejad et al., 2023; Zhou et al., 2024; Huang et al., 2025).

Training positional robustness. Randomized positional encodings expose training to sampled longer-range positions (Ruoss et al., 2023); FoPE and Resonance RoPE modify rotary features (Hua et al., 2025; Wang et al., 2024). Dropout trains under feature removal (Srivastava et al., 2014; Wager et al., 2013), while DropPos masks patch positions for reconstruction (Wang et al., 2023). Here the basis and span protocol stay fixed; complete harmonic pairs are masked under the copying objective. The interventions test learned use of that basis, and the theorem states sufficient conditions for margin protection. Expanded comparisons appear in Appendix M.

8 Discussion and conclusion

Training determines which comparisons a positional code supports. Different mixtures satisfy local demands while their distant responses interfere with context construction. Short-range forecasts locate consequential distant mass; fixed-basis dropout and signed transfers connect learned comparisons to context discrimination. Band interventions reveal useful alternatives, and the theorem states conditions for margin protection. Spectral power and short-copying success alone therefore do not determine a model's reach.

Pretrained-model support. In two sinusoidal XLNet QA checkpoints, short-context queries predict positional-intervention score changes at 4K tokens ($\rho = 0.840/0.955$). Editing source-name rows in a short-selected head raises one checkpoint's accuracy from 78.1% to 90.6%; the other gains margin but loses accuracy. Separate attention/value restorations reveal contributions from contextual representations (Appendix O).

Scope. The main evidence concerns small models and correct-history copying; context matching does not ensure occurrence selection or autonomous success. Forecast-selected edits test isolated outputs; equal-mass controls do not single out harmful addresses, and cohort associations with free-running L240 copying remain weak. The gap probe does not separate neural pathways; attention and accuracy remain distinct endpoints. Adverse grids and limited-band costs constrain dropout. The conditional theorem's native audit requires long-context mean-margin calibration, not a short-only guarantee. Pretrained probes neither predict complete QA behavior nor isolate dropout-mediated compensation.

AI USE STATEMENT

Generative AI tools assisted with developing hypotheses and conceptual models, designing experiments, implementing methods and synthetic-task generators, executing experiments, analyzing and interpreting results, formulating mathematical claims and proofs, preparing figures, and drafting and editing the manuscript. The authors are responsible for the final content of this work, including its text, code, claims and experimental results.

ETHICS STATEMENT

This work studies positional representations and length extrapolation using synthetic copying tasks, small transformers, and controlled span-QA probes of public pretrained checkpoints. Our generated evaluation tasks involve no human participants or collected personal data. Understanding failures beyond the training length may help assess model reliability, but the reported gains are specific to controlled tasks and do not establish reliable behavior in deployed systems. In particular, accuracy with correct output histories does not guarantee autonomous generation, and improved context matching does not resolve all repeated-occurrence ambiguities.

REPRODUCIBILITY STATEMENT

Section 2 specifies the task, architecture and evaluation measures. Appendix A defines the spectral readouts, frequency grids and intervention measurements; Appendix G records training budgets, optimizer settings, positional-span sampling, dropout normalization, seed pairing, cohort overlap and uncertainty estimates. The subset-margin theorem's proof and native measurement scope appear in Appendix J. Appendix O specifies the pretrained checkpoints, short-context selection and intervention controls. The experiments, analyses and figures can be reproduced using the code, configurations and saved results in the project repository.

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7aae9e3ec211249e05bd07271a6b1441-Abstract-Conference.html.

Krzysztof Choromanski, Shanda Li, Valerii Likhosherstov, Kumar Avinava Dubey, Shengjie Luo, Di He, Yiming Yang, Tamas Sarlos, Thomas Weingarten, and Adrian Weller. Learning a Fourier transform for linear relative positional encodings in transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2278–2286, 2024. URL https://proceedings.mlr.press/v238/choromanski24a.html.

Francesco D'Angelo, Oğuz Kaan Yüksel, Swathi Shree Narashiman, and Nicolas Flammarion. Induction heads interpolate N-grams. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2607.02800.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24932–24949, 2025. URL https://proceedings.mlr.press/v267/hua25b.html.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/928170bcb050fe64a63fe781b82265aa-Abstract-Conference.html.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21502–21521, 2024. URL https://proceedings.mlr.press/v235/jelassi24a.html.

Ryotaro Kawata, Yujin Song, Alberto Bietti, Naoki Nishikawa, Taiji Suzuki, Samuel Vaiter, and Denny Wu. From shortcut to induction head: How data diversity shapes algorithm selection in transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6499b639e8a4b5c9a780d9b88c09722f-Paper-Conference.pdf.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.19466.

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02795.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38018–38070, 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

Yui Oka, Kentaro Hanafusa, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Probing rotary position embeddings through frequency entropy. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0aee38a6fe9fffc8b658cfb1d872c1d5-Abstract-Conference.html.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency bands in RoPE: Base frequency and context length shape the interpolation–extrapolation trade-off. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/993dbaec0418a6449090f1debbcb8844-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish,

and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://papers.nips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1889–1903. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161/.

Jiajun Song, Zhuoyan Xu, and Yiqiao Zhong. Out-of-distribution generalization via composition: a lens through induction heads in transformers. *Proceedings of the National Academy of Sciences*, 122(6):e2417182122, 2025. doi: 10.1073/pnas.2417182122. URL https://doi.org/10.1073/pnas.2417182122.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://jmlr.org/papers/v15/srivastava14a.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, 2013. URL https://arxiv.org/abs/1307.1493.

Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and Zhaoxiang Zhang. DropPos: Pre-training vision transformers by reconstructing dropped positions. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2309.03576.

Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance RoPE: Improving context length generalization of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.32/.

Haodong Wen, Yiran Zhang, Yingfa Chen, and Kaifeng Lyu. Frontier language models struggle to copy: Text can be better viewed in 2D. *arXiv preprint arXiv:2607.16072*, 2026. URL https://arxiv.org/abs/2607.16072. Accepted at the FoGen Workshop, ICML 2026 (non-archival).

Xinyi Wu, Siyuan Liu, and Ali Jadbabaie. How data shapes RoPE frequency usage: From positional scale matching to length generalization. *arXiv preprint arXiv:2607.07678*, 2026. URL https://arxiv.org/abs/2607.07678. Accepted at the FoGen Workshop, ICML 2026 (non-archival).

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Joshua Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/45ed1a72597594c097152ef9cc187762-Abstract-Conference.html.

A MEASUREMENT DEFINITIONS AND SPECTRAL PREDICTION PROTOCOLS

Three spectral readouts. Learned-APE power is measured by a discrete Fourier transform over the position table. Early growth is the log power ratio between updates 512 and 1,536. The length experiment uses a 512-position cycle, vocabulary four, and four new seeds per maximum training length. Predictions were saved before these runs. The primary readout is first-layer signed cosine score demand at update 512, evaluated through profile ranks, exact maxima and adjacent-band directions; the subsequent power growth is a secondary readout. Exact individual maxima agree in 7/16 seeds for demand and 11/16 for growth, while all four group-growth maxima agree. Full-profile agreement is reported separately in Appendix B; exact-maximum agreement alone does not summarize the primary analysis.

The compensation and uncertainty studies instead use effective signed Fourier coefficients of the first-layer positional score before input-dependent normalization. Its displacement component has the form $K(d) = \sum_b [c_b \cos(\omega_b d) + s_b \sin(\omega_b d)]$; the associated band amplitude is $\sqrt{c_b^2 + s_b^2}$. This diagnostic is distinct from embedding power and from a complete native attention score. Fourier-feature inner products can approximate stationary kernels (Rahimi & Recht, 2007); our signed Q/K diagnostic can contain sine terms and is not assumed to be a symmetric positive-semidefinite kernel. Twelve learned-APE seeds at each of three sparsity settings give 36 models. The uncertainty model predicts each seed from the other eleven using a regularized Student-t distribution. Log-amplitude and signed-coefficient models are calibrated separately. Their marginal intervals and coverage audits are retrospective; they are not simultaneous guarantees over a spectrum.

Native utility and context advantage. Physical-band deletion replaces that band's coordinates by their positional means, without restoring RMS, and recomputes the complete network. Signed utility is the decrease in the specified margin. It may be negative and depends on the remaining bands. Conditional addition measures the reverse intervention under sampled partial supports. Neither utility nor its positive-part effective spread is identified with the signed Fourier coefficients. RoPE frequency-utilization diagnostics (Oka et al., 2026a) concern a different observable: our utility is a signed intervention effect on a specified comparison and can change with the competing address.

For a fixed correct/rival pair, M_m denotes correct-minus-rival native head score when m preceding tokens match. Its context advantage is $C = M_0 - M_{16}$. Planted-prefix probes preserve the current candidate tokens and the correct query context, and update the copied sequence consistently. Native head scores are pre-softmax quantities; their difference equals the log attention ratio for those two addresses. Constructed probes test conditional discrimination; ordinary copying accuracy uses the original evaluation distribution.

Vocabulary prediction. Four vocabulary-four seeds supply the reference spectrum, and eight disjoint seed identifiers per target vocabulary evaluate its dilation. The target profiles are signed context-advantage deletion utilities at $L = 240$, smoothed by one log-frequency spacing and normalized to unit Euclidean norm. Shape error is squared Euclidean distance. Reference and target seeds are bootstrapped together. Native context responses are a separate probe; broad spectral improvement does not imply that a single dilation predicts every context-response increment. All outcomes existed before the retrospective split was specified.

Frequency grids. Standard4, Standard8 and Standard16 use the indicated number of log-spaced frequencies between $10^{-3.75}$ and 1 radians per token. Base256-16 uses 16 log-spaced frequencies between $256^{-15/16}$ and 1. Geometric8 and Geometric16 span $2\pi/512$ to $\pi - 2\pi/512$ with the indicated log-spaced count. These labels identify fixed grids; changing the count preserves model width. Training and continuation budgets, mask normalization and cohort overlap are specified in Appendix G.

B SPECTRAL ACQUISITION AT FOUR TRAINING LENGTHS

The four training-length cases underlying Equation 2 compare the task-only prediction with measured early spectral growth. The predicted and measured mean peaks are bands 6, 4, 3 and 2 for maximum source lengths 60, 90, 120 and 180, respectively.

The plot is retained in the main text as Figure 2. At $T = 120$, the contrast gives $D_3 = 1.563 > D_4 = 1.126$ and $D_7 = 1.184 > D_6 = 0.929$. On a fixed 32-frequency sinusoidal basis, it selects $\omega = 0.03162$ (period 198.69), matching the early first-layer score-demand maximum in all four seeds. This auxiliary frequency count is distinct from the matched-width 4/8/16-band sweep in the main text.

Full-profile agreement is distinct from exact-peak agreement. The early-demand readout measures the negative token-loss derivative for adding a cosine score component centered at lag 1 to both first-layer heads at update 512. The original primary analysis includes rank correlation and adjacent-band direction agreement as well as the exact maximum. The secondary readout is log embedding Fourier-power growth from updates 512 to 1,536. Table 1 reports shape agreement over all 64 evaluated bands, without smoothing or selecting a favorable frequency window. Pearson correlations compare seed-mean profiles; direction agreement and the rank correlations below are computed within each seed and then averaged.

Table 1: **The task contrast captures spectral structure beyond the maximum.** Four seeds per length. Demand Pearson and adjacent-direction statistics are from the archived analysis; growth Pearson is a descriptive calculation from the same saved full profiles in this narrative revision.

Maximum length	Demand Pearson	Growth Pearson	Demand direction agreement
60	0.632	0.719	85.3%
90	0.799	0.749	86.9%
120	0.520	0.811	72.2%
180	0.559	0.815	60.3%

Across individual seeds, mean demand Spearman correlations are 0.573/0.532/0.396/0.126 at the four lengths; growth correlations are 0.468/0.391/0.449/0.160. Thus broad seed-mean agreement does not imply uniformly accurate fine-grained ranks in every run. A slow-first baseline gives growth rank correlations 0.475/0.400/0.227/0.125, slightly exceeding the task contrast at the two shorter lengths. Uniform-length weighting also gives higher mean-demand Pearson correlations at the first three lengths (0.909/0.845/0.544, versus 0.120 at length 180). The results support structured, length-dependent preferences without establishing that token-exposure weighting is uniquely correct or that it wins every profile metric.

C Compensation and uncertainty across learned spectra

Task demands constrain the comparisons that bands jointly produce. Independent measurements show how different band mixtures can compensate. Across 36 learned APE models, final signed Fourier-coefficient variation is **67–105%**, versus **11–22%** for short-range positional comparisons (relative RMS; Figure 7). Retaining between-band covariance reduces comparison variance by 73–86% across the two measured checkpoints; an amplitude-parallel projection retains a 71–81% reduction. Substantial spectral variation can therefore coexist with stable comparisons through cancellation between bands.

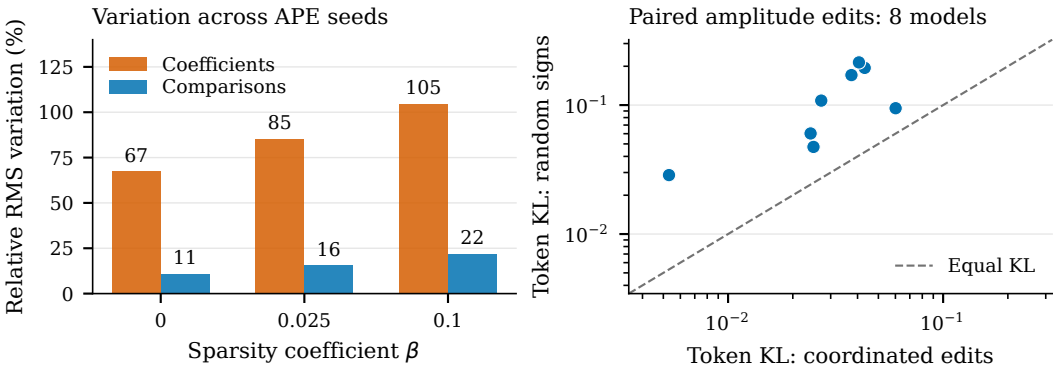


Figure 7: **Bands compensate in measured comparisons and controlled edits.** Left: final coefficient and comparison variation, twelve APE seeds per sparsity setting. Comparisons use the first-layer positional kernel before input-dependent normalization, between lag 1 and lags 0–120. Right: held-out token KL for eight sinusoidal models under coordinated amplitude edits and randomized-sign controls, at 15% perturbation with positional RMS restored. Each point is one model; the line marks equality. Perturbation magnitudes match before RMS restoration.

Compensation also preserves native behavior approximately. Coordinated amplitude edits selected on separate calibration examples produce token KL of **0.0329**, versus **0.1148** for randomized-sign controls; accuracy losses are **0.98 versus 3.05 percentage points**. Margin RMS changes are 0.907 versus 1.548; both margin and KL disruption are smaller in all eight models. These interventions demonstrate functional alternatives among active bands. This compensation explains why retaining correlations narrows the uncertainty over comparisons in Figure 8.

Uncertainty over acquired bands. Across seeds, the learned spectrum varies. Retrospective leave-one-seed-out predictions give about 90% coverage for band amplitudes and 93% for positional comparisons with nominal 90% intervals (Figure 8). Retaining correlations between bands makes comparison intervals 43–55% narrower than treating bands independently. The compensation measurements above explain how different band mixtures can preserve these comparisons.

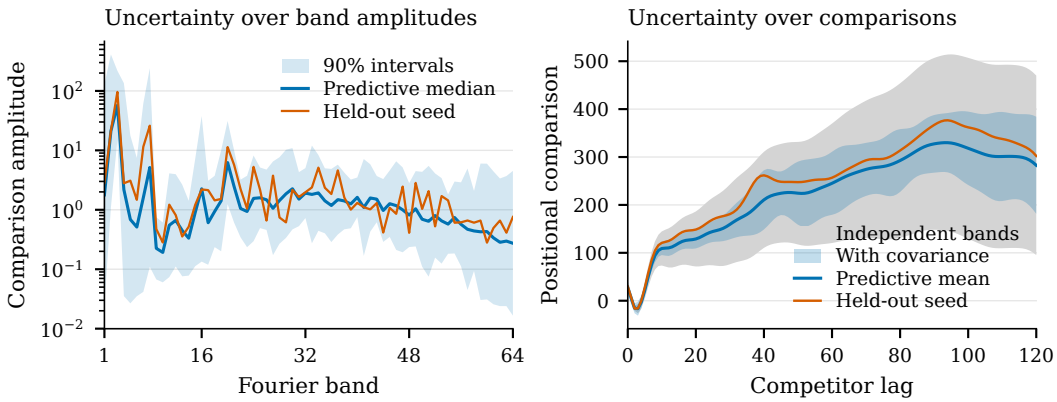


Figure 8: **Predicting a family of spectra and its shared comparisons.** An APE seed predicted from eleven others: marginal 90% amplitude intervals (left) and comparison intervals (right). Separate amplitude and signed-coefficient models use first-layer positional diagnostics, calibrated independently of the task-only prediction in Equation 2.

D ADDITIONAL COMPARISON AND INTERVENTION DETAILS

Band contributions depend on the task comparison. In the paired 64-token context test, deleting $\omega = 0.005623$ reduces ordinary matched-context margin by 1.322 units and extra context discrimination by 2.871. At $\omega = 0.177828$, the effects are 0.239 and 15.091 units. Bands thus have overlapping roles with different weights.

A useful band need not ensure correct occurrence selection. On the planted two-address challenge, preferences oscillate with positional gap, but correct ranking averages only 51.9% ordinary and 48.8% dropout. Deleting $\omega = 0.03162$ lowers ordinary correct ranking to 31.3%. Signed preference and token correctness remain distinct from unsigned separation. Joint positional-band removal reduces context-response RMS from 32.43 to 1.06 in ordinary models and from 28.86 to 0.35 with dropout.

Two independent geometry controls. The PE-gap experiment keeps the full token sequence fixed and translates the separator and copied segment by 0–512 positional coordinates, in steps of 16. The source candidates remain 96 tokens apart and share 64 preceding tokens. Its gap-dependent changes measure a positional intervention with content fixed. The balanced physical-displacement experiment instead places rivals on both sides of the correct source and crosses displacement with 0/4/8/16 matching preceding tokens. It uses 16 backgrounds and common origins 0/30/60 at each source length. Across the matched/mismatched arms at a given displacement, addresses and candidate tokens are fixed. These protocols measure different positional changes and are not pooled.

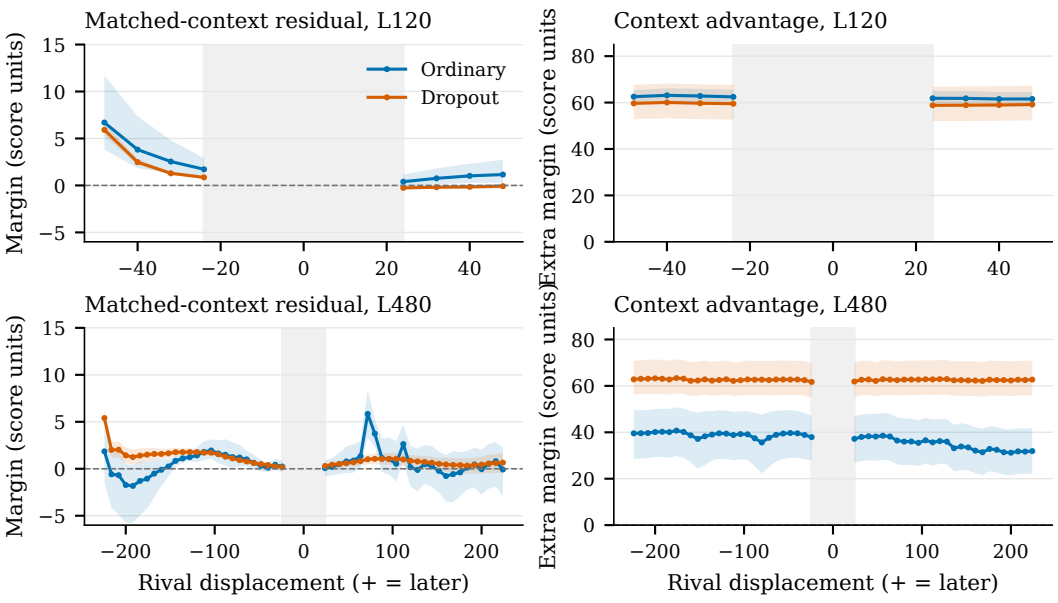


Figure 9: **Context mismatch supplies a large advantage across rival displacements.** Left: correct-minus-rival scores with 16 matching preceding tokens. Right: the extra margin when that rival's context mismatches, at the same physical displacement. Gray marks the unmeasured central interval. Twelve seeds per recipe; pointwise 95% seed-bootstrap intervals. These conditional score curves complement the fixed-token PE-gap intervention in Figure 1.

A gap-induced change identifies the effect of the positional intervention. The absolute score at one gap can additionally use longer history or landmarks; the sweep does not require removing them. Context and position can interact inside the network, so the controlled effects are not assumed to be additive neural pathways. The balanced displacement curves retain this interaction by reporting the context contrast separately at each displacement.

E REPEATED CONTEXTS EXPOSE ADDRESS AMBIGUITY

Repeated n-grams limit context-based copying (Jelassi et al., 2024; Wen et al., 2026). Our probes separate token correctness from intended-address selection: exact repetition can hide addressing errors because several addresses supply the same token. At source length 120 and repeat period 16, both training recipes achieve 99.8% accuracy on exact repetitions. Varying the payload after each shared 15-token prefix lowers accuracy to **56.8% ordinary and 55.7% dropout**. Occurrence ambiguity matters even within the training range; payload edits also change contextual evidence.

Dropout preserves useful information beyond the common prefix. With K repeated prefixes queried equally, a predictor given the source and only the common prefix as query information must give the same prediction at each occurrence. Its token accuracy is bounded by

$$B(x) = \frac{\max_y n_y}{K}, \tag{5}$$

where n_y counts occurrences followed by token y. For 15 occurrences sharing seven tokens at $L = 480$, ordinary/dropout accuracy is **42.3%/62.6% against a 39.2% bound**. Dropout's excess is 23.4 percentage points (95% CI [20.6, 26.3]), with all twelve seeds above the bound. Correct address ranking also improves from 21.2% to 33.3%. At a shared prefix of 15 tokens, its excess remains 11.3 points [8.2, 14.3]. These results establish useful occurrence-dependent information beyond the shared suffix, potentially supplied by longer history, position, or landmarks.

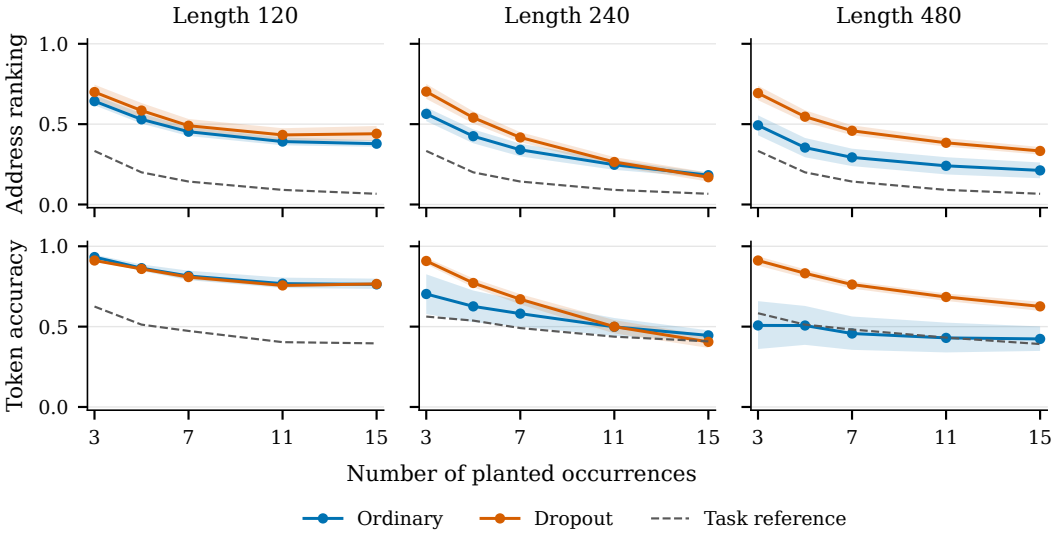


Figure 10: **Intervention benefits depend on the remaining occurrence ambiguity.** Seven-token shared prefixes, spacing eight, all planted occurrences queried equally. Top: correct ranking among the K addresses, with reference $1/K$. Bottom: model token accuracy, with shared-prefix-only upper bound $B(x)$. Address metrics are averaged over heads within each model; token accuracy is measured at the model output. Twelve seeds per recipe; shading gives pointwise 95% seed-bootstrap intervals.

At $L = 240, K = 15$, dropout instead reaches 40.5% against a 40.8% bound. Useful information beyond the common prefix depends on the configuration; improved context selection need not resolve every occurrence.

F REPEATED-ADDRESS ERRORS HAVE STRUCTURED DESTINATIONS

The repetition-count comparison measures how discrimination changes as more occurrences become plausible. The confusion distributions below additionally show where incorrect selections go. At $L = 240$, dropout often selects the final candidate across different targets; at $L = 480$, its errors are more concentrated near the correct occurrence. This configuration dependence complements the token-accuracy and signed-margin measurements.

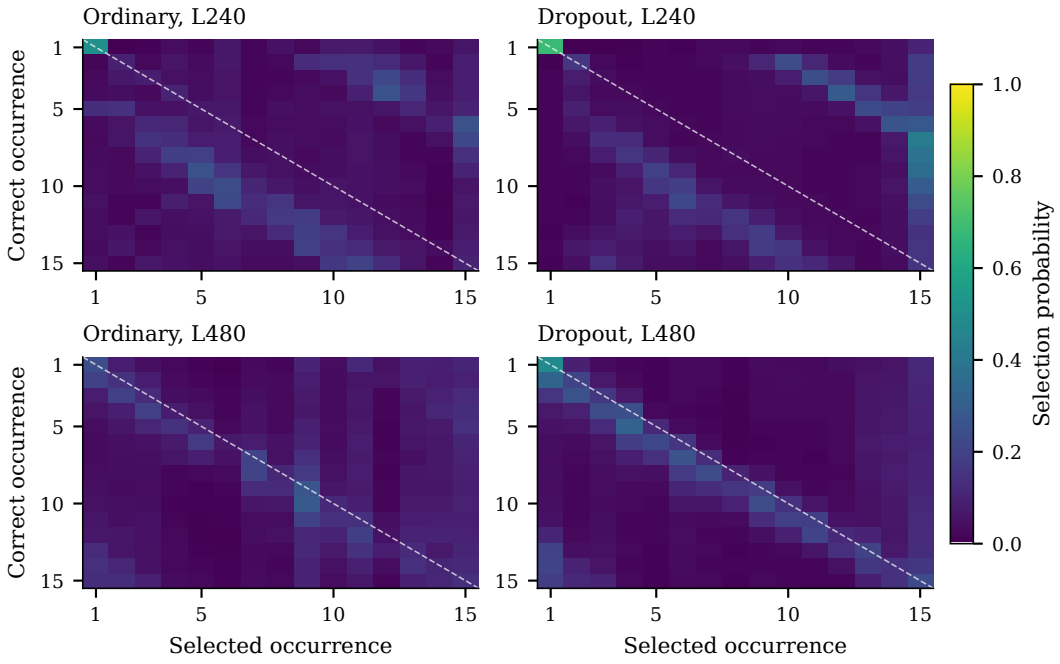


Figure 11: Correct versus selected occurrence among fifteen candidates sharing a 15-token prefix, spaced sixteen tokens apart. Each row is a probability distribution; the diagonal records correct selection. Probabilities average native head winners over backgrounds, origins and twelve seeds; these are address choices, including wrong addresses with the correct payload.

G COMPLETE TRAINING-INTERVENTION COHORTS

Six-grid additive comparison. The completed 216-model study uses vocabulary four, six grids, $p = 0, .25, .5$ and twelve paired seeds per cell. Every model has width 64, two heads and two blocks; all train for 65,536 Adam updates with batch size 128 and learning rate 10^{-3}, on source lengths through 120. Random positional spans are nonwrapping in a 1,024-position domain. Ordinary and dropout arms share initialization and task/origin sampling. A mask retains each complete sine/cosine component with probability $1 - p$; if none survives, one is retained uniformly. With K supplied and M retained components, amplitudes are scaled by $\sqrt{K/M}$. Evaluation restores every component.

Table 2: All six grids and both doses. Entries are dropout-minus-ordinary accuracy differences in percentage points. Short is $L = 120$; long is $L = 480$. TF uses correct output histories. The new-ten column excludes the two developmental seeds and overlaps the twelve-seed estimate. Greedy differences are descriptive means, not exact-copy rates.

Basis	p	Short TF	Long TF [95% CI]	New ten TF	Long greedy
Standard4	0.25	-0.21	-4.43 [-29.11, +20.25]	-5.97	-4.58
Standard4	0.50	-1.44	-1.80 [-22.32, +18.71]	+1.99	-6.40
Standard8	0.25	-0.02	+41.24 [+33.99, +48.48]	+41.81	+2.17
Standard8	0.50	+0.06	+35.59 [+28.47, +42.72]	+38.88	+1.08
Standard16	0.25	-0.06	+33.98 [+18.04, +49.91]	+36.61	+36.95
Standard16	0.50	-0.28	+30.00 [+13.86, +46.14]	+32.29	+11.29
Base256-16	0.25	-0.01	+35.48 [+27.73, +43.23]	+35.50	+7.18
Base256-16	0.50	-4.65	+39.12 [+29.33, +48.91]	+40.12	+4.68
Geometric8	0.25	-3.71	-15.50 [-20.44, -10.56]	-14.87	-1.81
Geometric8	0.50	-20.82	-27.02 [-34.65, -19.40]	-28.41	-2.09
Geometric16	0.25	-0.08	+37.38 [+31.30, +43.47]	+39.63	+2.82
Geometric16	0.50	-4.13	+32.62 [+18.51, +46.73]	+29.38	+1.79

All acquisition outcomes are included. Intervals are paired 95% Student-t intervals over training seeds, without multiplicity adjustment. Tokens, heads and evaluation positions are not independent experimental replicates. Higher dropout doses can reduce short accuracy: the 50% means on Base256-16 and Geometric16 fall by 4.65 and 4.13 points, respectively. On Geometric8, 50% dropout lowers short accuracy by 20.82 points as well as harming long accuracy. The four favorable 25% contrasts instead retain mean short accuracy within 0.1 points of ordinary training.

Basis and vocabulary coverage. The 360-condition study crosses Standard16 and Geometric8 in additive and rotary form, plus a custom four-plane additive code, with vocabularies 2/4/8, ordinary versus selected component dropout, and twelve seeds. The selected dose is 25% for Sinusoidal/RoPE and 50% for custom planes. The architecture and training budget match the six-grid study. Two seeds informed developmental choices; ten additional seeds provide confirmation.

The Standard16/Geometric8 additive V4 ordinary and 25% cells overlap the six-grid study and are reused. Counts of 216 and 360 describe cohort designs, not disjoint model sets. The original larger 864-condition plan is not a completed factorial and is not used as the evidence count.

Table 3: Complete five-configuration, three-vocabulary comparison. Teacher-forced content accuracy (%) and paired long-length gain (percentage points), all twelve seeds including acquisition failures. Intervals are pointwise paired 95% Student-t intervals.

Family / basis	V	$L = 120$ Ord./Drop	$L = 480$ Ord./Drop	Gain [95% CI]
Sin. Standard16	2	93.11/78.01	48.17/57.02	+8.84 [+2.47, +15.22]
Sin. Standard16	4	99.95/99.89	64.84/98.81	+33.98 [+18.04, +49.91]
Sin. Standard16	8	99.98/99.99	83.96/99.95	+15.99 [+5.58, +26.40]
Custom four-plane	2	96.75/85.46	48.82/49.51	+0.69 [-12.86, +14.23]
Custom four-plane	4	99.48/99.99	57.48/89.59	+32.12 [+16.09, +48.14]
Custom four-plane	8	97.79/99.06	23.72/69.78	+46.06 [+38.30, +53.83]
Sin. Geometric8	2	75.46/71.38	46.34/41.49	-4.85 [-14.87, +5.18]
Sin. Geometric8	4	99.01/95.29	43.49/27.99	-15.50 [-20.44, -10.56]
Sin. Geometric8	8	99.93/99.81	59.62/42.35	-17.27 [-19.91, -14.62]
RoPE Standard16	2	92.77/98.37	50.31/44.82	-5.48 [-10.35, -0.62]
RoPE Standard16	4	99.98/99.48	25.58/24.97	-0.60 [-1.46, +0.25]
RoPE Standard16	8	99.99/99.92	13.04/17.29	+4.25 [+2.53, +5.97]
RoPE Geometric8	2	87.33/96.40	49.45/45.36	-4.09 [-8.02, -0.17]
RoPE Geometric8	4	94.16/97.72	24.10/23.98	-0.11 [-1.77, +1.54]
RoPE Geometric8	8	99.95/99.12	11.82/12.53	+0.71 [-0.81, +2.23]

Sinusoidal/RoPE use nonwrapping spans and no generic layer dropout. The custom code uses frequencies $2\pi\{20, 37, 54, 83\}/512$, cyclic origins and retained 0.1 attention/MLP dropout. Its $L = 480$ evaluation includes wraparound stress. Within-basis ordinary/dropout contrasts are matched; absolute cross-family accuracy is not an isolated encoding-family effect. RoPE masking removes Q/K content coordinates as well as their rotary contributions.

For Standard16 at V4/V8, the ten confirmation seeds retain long-length gains of 36.61/15.47 points; custom-code gains are 30.45/46.12 points. Vocabulary two can show improved long accuracy while losing short acquisition, as in Standard16 (93.11% to 78.01% short accuracy). At V8, RoPE Standard16 improves by 4.25 points but remains at only 17.29% long accuracy. These results support benefits across several additive bases and vocabularies, not a universal rescue of positional encodings.

Continuation and execution cohorts. The main-text matched Standard16/Geometric8 continuations use separate V4 endpoints: each 8,192-update ordinary parent continues for 8,192 updates in each arm. Their accuracy is not the 65,536-update endpoint in the tables above. A further acquired Standard16 V8 cohort has twelve matched parents and the same 8,192-update continuation budget; 25% dropout raises $L = 480$ accuracy from 91.85% to 99.58%, a gain of 7.72 points [3.14, 12.31]. The historical generic-corruption comparator missed its intended short-loss match, so it does not isolate frequency compensation from other regularization effects.

The autonomous custom-code confirmation is also separate: twelve seeds per arm, V4, and 2,048 IID sources per model at $L = 240$. Stopped accuracy counts content positions after early EOS as wrong; exact copy with EOS requires the entire correct copy and termination. Ordinary versus normalized dropout gives 71.16/99.65% stopped accuracy and 0/76.35% exact copy with EOS. These outcomes support autonomous execution on that cohort; they do not replace the teacher-forced endpoints or establish autonomous success on every grid.

Relation to the task partition. Selecting a matching-context class and resolving its correct occurrence remain distinct requirements. Dropout improves the former on the planted two-address challenge; repetition-count tests show useful information beyond the common prefix at long length. Signed transfers identify a contributing positional comparison, without isolating the full path from masking through either task requirement to the cohort-wide behavioral gains.

H ADDITIONAL REPAIR AND TRANSFER RESULTS

Preserving context comparisons improves extrapolation. The local comparison in Section 5 provides an intervention target. Fitting first-layer attention to preserve its clean short-input distribution under component removal improves eight-band L480 accuracy by 2.32–3.11 points, with nearly unchanged short copying and no long-input fitting (Appendix L). In a separate twelve-seed Standard16 cohort, 512 paired clean/masked updates give **64.5% versus 86.1%** long token accuracy. Transferring only the learned near–far content competition reaches **82.9%**; restoring old competition lowers the masked model to 70.1%. Recipient values, marker weights and within-group attention remain fixed. Query construction carries the larger effect, with additional source and joint contributions (Appendix L).

These changes depend on the available bands. In the 4/8/16-band comparison, dropout raises the lag-1-versus-distant margin, while short context-class selection changes by -14.4 **points with four bands and** $+17.1$ **with sixteen**. A limited basis can incur a local-discrimination cost; richer supports improve both in this experiment (Figure 14).

Extrapolation improves after short copying is acquired. Matched continuations test repair after acquisition with the basis and training time fixed. In a separate vocabulary-four Standard16 cohort, dropout raises $L = 480$ accuracy from **71.18% to 97.73%**, with mean $L = 120$ accuracy of **99.82% in both arms** (Table 4). The paired gain is 26.55 percentage points (95% CI [15.35, 38.02]).

Table 4: **Equal training time, different extrapolation.** Accuracy (%) under correct histories. Same ordinary parent, 8,192 additional updates per arm, 25% dropout; twelve pairs per basis, maximum training length 120.

	$L = 120$		$L = 480$	
Basis	Ordinary	Dropout	Ordinary	Dropout
Standard sinusoidal (16 bands)	99.82	99.82	71.18	97.73
Geometric (8 bands)	95.13	93.29	38.32	26.63

Signed positional comparisons carry a transferable part of the gain. We transfer signed sine/cosine coefficients of the first-layer displacement-dependent positional QK numerator between paired models. The recipient stays fixed, with native normalization active; transfer size is selected and independently validated on short examples.

Moving an ordinary model toward dropout improves long-length accuracy by **3.13 points** (95% CI [1.50, 4.92]); moving away changes it by -8.71 **points** ([-11.99, -5.94]). Transferring the position-sum coefficients instead has approximately zero mean effect (Figure 5). The effect is partial and direction-sensitive.

Transfer sizes differ across directions and controls. Both directions can hurt the dropout recipient, consistent with locally co-adapted comparisons.

The adverse basis also transfers an adverse comparison. On Geometric8, dropout lowers long-length accuracy from 38.32% to 26.63% and short accuracy from 95.13% to 93.29%. Moving an ordinary recipient toward dropout likewise lowers long accuracy by 4.20 points [2.91, 5.59]; 9/12 transfers pass short validation. The transfer direction tracks the training effect in both bases. We next examine which changes in band use accompany stronger long-range comparisons.

I STANDARD16 COMPARISON DIAGNOSTICS

The following diagnostics examine Standard16 within the broader training comparison. At four times the maximum training length, next-token error falls from **35.16% to 1.20% for V4** and **16.04% to 0.05% for V8**, with improvement in all twelve seeds of each vocabulary (Figure 12). Correct output histories exclude generated error cascades from this comparison.

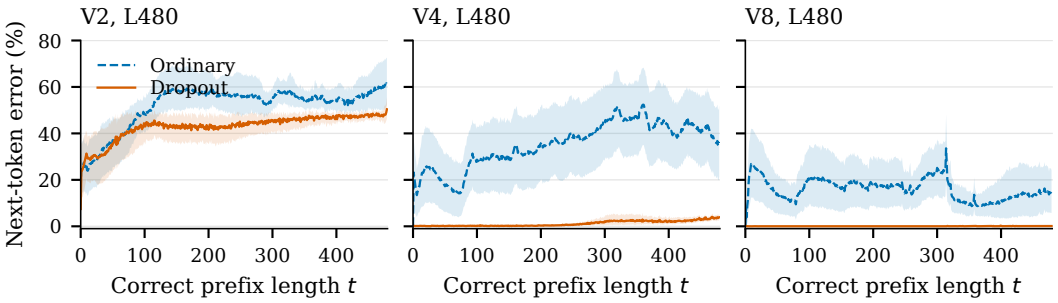


Figure 12: **Dropout preserves long-length prediction across output positions.** Error on the next copied token after t correct output tokens, at source length 480; maximum training length 120, component dropout 25%. Twelve paired seeds per vocabulary, all included; shading is pointwise 95% uncertainty in the seed mean. All positional components are present at evaluation.

The learned comparisons become less vulnerable to individual bands. For ordinary models, the largest mean deletion effect on correct-versus-strongest-rival margin shifts from $\omega = 0.1778$ at $L = 120$ (2.334 score units) to $\omega = 0.003162$ at $L = 480$ (4.023 units). Dropout substantially reduces these single-band effects (Figure 13). Together with the joint-deletion collapse of context discrimination, this shows robust use of band combinations: small individual effects coexist with strong collective dependence on the positional code.

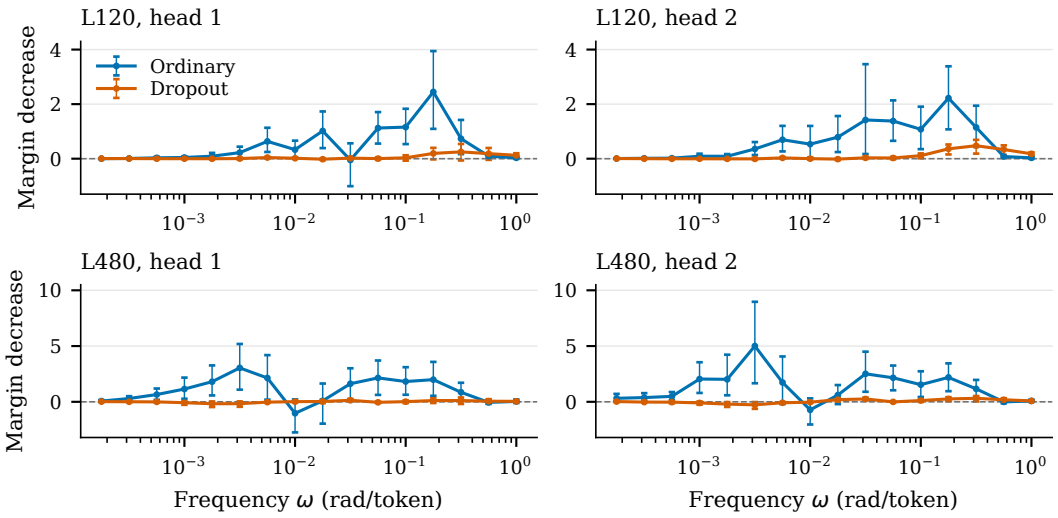


Figure 13: **Long-length addressing exposes a different pattern of band reliance.** Signed mean margin decrease after removing one physical frequency band, shown separately for both heads. The strongest incorrect source is recomputed after mean-preserving deletion, without norm restoration. Twelve Sinusoidal16 seeds per recipe; pointwise 95% seed-bootstrap intervals.

Task difficulty also affects acquisition. For V2, long-length error decreases from 51.88% to 42.98%, but only 3/12 dropout seeds reach 95% training-length accuracy, compared with 10/12 ordinary seeds.

J SUBSET ROBUSTNESS AND COMPLEMENTARY SEPARATION

Write $\phi_i(d) = 1 - \cos(\omega_i d)$, $M_S(d) = \sum_{i \in S} a_i \phi_i(d)$, with nonnegative coefficients of fixed total mass. For a uniform k-subset, independent of a short-range sample t, assume $\mathbb{E}[(\gamma - M_S(t))_+^2] \le \varepsilon^2$. At any $d \in E$, choose g bands G with $\phi_i(d) \ge \kappa$. Let $H = \{S \subseteq G\}$, which has probability $q = \binom{g}{k}/\binom{n}{k}$. Cauchy–Schwarz gives

$$\mathbb{E}[M_S(t) \mid H] \ge (\gamma - \varepsilon/\sqrt{q})_+ .$$

Since $\phi_i(t) \le 2$ and each $i \in G$ appears with conditional probability k/g, the same expectation is at most $(2k/g) \sum_{i \in G} a_i$. Finally $M_{\mathrm{full}}(d) \ge \kappa \sum_{i \in G} a_i$, proving the main bound uniformly on E. The assumption concerns the distribution over masks, not only a finite observed training sample. It does not assert optimizer convergence.

Why short full-support success is insufficient. For frequencies $(\pi, \pi/2)$, training displacement 1 and distant displacement 2, take a fixed weight budget $3\gamma/2$. Ordinary short fitting admits $(a_1, a_2) = (3\gamma/2, 0)$, with short margin 3γ and distant margin zero. Requiring each singleton to retain margin γ instead forces $(a_1, a_2) = (\gamma/2, \gamma)$, giving distant margin 2γ. This illustrates a stronger feasible-class guarantee, not pointwise superiority of every dropout model over every ordinary model.

Mask scale. The main theorem uses unscaled retained contributions. For a constant masked scale $\alpha = n/k$, divide its positive bound by α when γ and ε refer to scaled masked margins. Native component dropout restores PE RMS, but neural margin scaling need not equal PE scaling. That effect belongs in an explicit representation or its residual.

J.1 SIGNED EXTENSION WITH SHORT- AND LONG-RANGE RESIDUALS

For the context contrast, suppose a shared representation across masks is

$$C_S(z) = b(z) + H_S(z) + r_S(z), \qquad H_S(z) = \sum_{i \in S} A_i f_i(z), \quad A_i \ge 0,$$

where $b = C_\varnothing$ and z includes the declared content and positional conditions. The functions retain phases and may be negative. For example, $f_i(d) = \cos\phi_i - \cos(\omega_i d + \phi_i)$ is a signed harmonic difference. Such a decomposition of the neural context advantage is an assumption.

Suppose the RMS shortfall below γ of $C_S - b$ is at most ε, and $\|[r_S(t)]_+\|_2 \le \eta_T$. Choose $\beta_i \ge \sup_{t \in T}[f_i(t)]_+$ and, for $\rho > 0$, define $G(z) = \{i : \beta_i > 0,\ f_i(z) \ge \rho\beta_i\}$, with $g(z) \ge k$ and $q(z) = \binom{g(z)}{k}/\binom{n}{k}$. Let $N(z) = \sum_i A_i[-f_i(z)]_+$ and assume $r_{\mathrm{full}}(z) \ge -\eta_E(z)$. Then

$$C_{\mathrm{full}}(z) \ge b(z) + \rho \frac{g(z)}{k} \left[\gamma - \frac{\varepsilon + \eta_T}{\sqrt{q(z)}}\right]_+ - N(z) - \eta_E(z).$$

The same conditional expectation argument applies to H_S: its lower bound pays for both shortfall and positive short residual, and its upper bound is $(k/g) \sum_{i \in G} A_i \beta_i$. At z, add the favorable contributions and subtract destructive terms and the adverse residual. Taking the infimum gives a uniform floor on the declared domain. Improving this certificate alone does not prove an increase in the measured minimum.

K MEASURED MARGIN RETENTION AND USEFUL SPREAD

For each seed, average context advantage over backgrounds, origins and heads, then take the minimum over eight displacement ratios shared across lengths. Divide the cohort mean of these minima at $L = 480$ by the cohort mean context advantage at $L = 120$. This is a sampled mean-curve statistic, not a worst-case bound over individual queries or all positions. Standard4 shows why retention must be read alongside absolute margin: a smaller short denominator can raise the ratio without raising the long minimum.

Table 5: All six bases and three doses: short mean, minimum sampled long curve, their ratio, and effective conditional utility spread. Twelve seeds per row, 216 models in total. Utility spread uses the separate compact intervention panel.

Basis / dropout	Short mean	Long minimum	Retained (%)	Useful spread
Standard4 / 0%	40.29	42.96	106.6	1.80
Standard4 / 25%	30.54	41.50	135.9	1.59
Standard4 / 50%	34.39	52.20	151.8	1.56
Standard8 / 0%	51.21	16.79	32.8	3.92
Standard8 / 25%	42.95	47.19	109.9	4.38
Standard8 / 50%	39.93	44.15	110.6	4.71
Standard16 / 0%	62.20	30.55	49.1	6.35
Standard16 / 25%	59.34	61.97	104.4	10.24
Standard16 / 50%	53.62	56.10	104.6	10.71
Base256-16 / 0%	66.33	-1.13	-1.7	5.19
Base256-16 / 25%	55.80	39.43	70.7	11.16
Base256-16 / 50%	34.18	28.43	83.2	10.46
Geometric8 / 0%	61.91	24.25	39.2	5.28
Geometric8 / 25%	24.42	5.32	21.8	5.44
Geometric8 / 50%	23.72	0.70	2.9	3.86
Geometric16 / 0%	70.09	11.21	16.0	7.20
Geometric16 / 25%	57.75	42.69	73.9	7.73
Geometric16 / 50%	31.29	23.10	73.8	8.58

Distinct probe panels. Scale matching uses short displacements $\pm 24, \pm 32, \pm 40, \pm 48$, multiplied by $L/120$ at longer lengths, with the original discrimination backgrounds/origins. The compact utility panel instead uses $L = 480$ displacements $\pm 32, \pm 96, \pm 160, \pm 224$, four backgrounds and two origins. Its Standard16 ordinary/dropout minima are 30.09/63.65, whereas the scale-matched minima are 30.55/61.97. Their normalizations are kept separate.

Utility and uncertainty. Physical-band removal replaces that band's coordinates by their positional means, without norm restoration. Native normalization and all layers are recomputed. Conditional utility averages addition effects over four shared Bernoulli-half-support masks. Utilities can be negative and nonadditive; effective spread uses only their positive parts, computed per seed. Bootstrap intervals resample twelve trained seeds, conditional on these probes. The mask diagnostic is not an exact additive allocation or a measured coefficient vector for the theorem.

Residual audit. For any common additive representation with uniform residual bound η, the range of a band's conditional utility across supports cannot exceed 4η. The measured ranges therefore imply necessary residual allowances of at least 6.08 ordinary and 4.88 dropout for Standard16's mean curves. These are lower bounds on required error, not the upper bounds η_T, η_E needed for a neural certificate. The separate audit below measures those upper bounds on a declared probe distribution; the earlier utility panel alone cannot supply them.

K.1 MEASURING THE NATIVE PREMISES AND RESIDUAL UPPER BOUNDS

We additionally audit all six V4 bases, ordinary and 25% dropout, with twelve paired seeds: 144 frozen width64, two-head, two-block models. This tests a *sufficient condition in native networks*; it does not identify unique causal mediation or establish prediction from short data alone.

Exhaustive short masks and a signed fit. We retain $k = 3n/4$ distinct physical bands, zero the other coordinates, and multiply retained PE coordinates by $\sqrt{n/k}$. This matches the training intervention conditional on retaining k bands. All $\binom{n}{k}$ masks are evaluated: 1820, 28, or 4 masks for 16, 8, or 4 bands. The short panel has 16 fresh backgrounds, independently sampled nonwrapping origins in $\{0, \dots, 60\}$, and displacements $\pm 24, \pm 32, \pm 40, \pm 48$ at $L = 120$. Both heads are averaged. Native subset competence means that every sampled displacement retains at least half the full-support short mean context margin. Its mask fraction is enumerated exactly, conditional on this finite panel.

Using only 64 sampled short masks and the full-support short scores, we fit shared signed functions

$$h_i(d) = \theta_{i0} + \theta_{ic} \cos(\omega_i d) + \theta_{is} \sin(\omega_i d).$$

The ridge coefficient is fixed at .01 after scaling targets by the short RMS margin; full-support observations carry 20% of the fitting loss. Constants are included because context advantage need not vanish at displacement zero. The score surrogate uses $\alpha = n/k$ on retained functions and $\alpha = 1$ at full support. Departures from this scaling, content effects and mask interactions belong to $r_S = C_S - \alpha \sum_{i \in S} h_i$; they are not discarded.

The native shortfall $\varepsilon(\gamma)$ and positive residual RMS η_T are computed on *every* mask and short displacement. Substituting them in the signed theorem, with $\beta_i = \max_t [h_i(t)]_+$, gives a harmonic lower bound $B(d)$ that pays for negative band contributions. We optimize the valid bound over a fixed grid of 257 positive margin thresholds and the finite coverage ratios $h_i(d)/\beta_i$. These choices use short observations and known harmonic features, without long neural observations.

An upper residual bound with a stated population. At $L = 480$, the eight displacements are multiplied by four. We draw 127 calibration panels and 128 test panels, each averaging 16 independent fresh background/origin draws and both heads. The fit is frozen before calibration. For a panel X, let

$$Z(X) = \max_{d \in E} \left[\sum_i h_i(d) - C_{\text{full}}(d; X) \right]_+ .$$

The 125th ordered calibration score $Z_{(125)}$ is a one-sided nonparametric tolerance bound: $\Pr\{\text{Bin}(127, .95) \geq 125\} = .04424$. Consequently, with at least 95% confidence over calibration, at least 95% of new panels from this distribution satisfy, simultaneously over the eight displacements,

$$C_{\text{full}}(d; X) \geq B(d) - Z_{(125)}.$$

The guarantee is per model and concerns panel means, not individual queries or all token strings. A retained four-model pilot calibrated individual-background residuals instead; its subset-theorem floors were nonpositive. The new panel claim must not be read as repairing that stronger failed guarantee.

Population mean certificate. We also derive a global score range from the final LayerNorm and Q/K weights. For centered normalized states x, $\|x\| \leq \sqrt{w}$, with $w = 64$. After absorbing LayerNorm gain and centering into Q, K, and its shift and query bias into q_0, put $A = Q^\top K/(2\sqrt{32})$ and $v = K^\top q_0/(2\sqrt{32})$. Key bias cancels in a target–rival difference. The two-head-averaged context contrast therefore obeys, for every input,

$$|C| \leq B_C = 4w\|A\|_{\text{op}} + 4\sqrt{w}\|v\|_2.$$

Apply the empirical Bernstein inequality of Maurer & Pontil (2009) to the 2032 calibration backgrounds at each displacement. If $\bar{C}_d, s_d^2$ are their mean and unbiased variance, define

$$u_d = \sqrt{\frac{2s_d^2 \log(2J/.05)}{2032}} + \frac{14B_C \log(2J/.05)}{3(2032 - 1)}.$$

Using $J = 8 \times 144$ gives a simultaneous 95% confidence statement over all declared model/displacement combinations, without assuming independence across models. The expected adverse residual is at most $\eta_E(d) = [\sum_i h_i(d) - \bar{C}_d + u_d]_+$, hence $\mathbb{E}C_{\text{full}}(d) \geq B(d) - \eta_E(d)$.

Table 6: Native certificate audit. Competence is the mean fraction of all retained subsets meeting the short threshold. Counts report models with strictly positive lower bounds. Panel certificates use the per-model 95% confidence/95% content tolerance bound. Mean certificates use the simultaneous 95% confidence bound over all 144 models and eight displacements. Each ordinary/dropout pair contains twelve seeds.

Basis	Competence (%) ord. / drop.	Signed panel ord. / drop.	Signed mean ord. / drop.	Phase-free panel ord. / drop.
Standard4	62.5 / 77.1	0 / 1	0 / 0	0 / 4
Standard8	75.9 / 99.7	0 / 11	0 / 3	0 / 12
Standard16	82.5 / 100.0	0 / 12	0 / 6	0 / 12
Base256-16	91.1 / 100.0	0 / 1	0 / 0	0 / 12
Geometric8	80.4 / 91.7	0 / 0	0 / 0	0 / 0
Geometric16	92.5 / 100.0	0 / 3	0 / 0	0 / 12

This additional mean analysis uses the fixed-size calibration data; it was introduced after the developmental pilot. Numerical evaluation includes a .001 score allowance and .002 allowance in the combined short errors.

Excluding the developmental seed leaves 66 ordinary and 66 dropout models; 0 and 26, respectively, retain positive signed panel certificates. All twelve seeds remain in the table. The shared calibration/test backgrounds are not extra trained-model replicates.

Using the same threshold for each pair—half the ordinary model's short full-support mean—gives competent-mask fractions of 82.5% versus 99.8% for Standard16, but 80.4% versus 8.3% for Geometric8. Thus a high fraction relative to a model's own weak short margin must not be mistaken for strong competence.

What the control permits us to conclude. A phase-free control fits one constant per band using the same short data, masks, ridge penalty and calibration procedure. A constant-short-mean comparator is also calibrated on the same long panels. Strong phase-free bounds show that successful calibration alone does not establish the unique predictive value of harmonic phase geometry. Long-context calibration remains essential here. The audit can verify native subset competence and nonvacuous, distribution-specific sufficient conditions; it cannot establish that short-only training guarantees extrapolation, that every query is protected, or that this decomposition uniquely mediates the dropout gain in copying accuracy.

Spread predicts part of the measured flattening. For similarly scaled, weakly correlated band fluctuations at a matched horizon, the normalized mean-to-minimum deficit D approximately scales as $n_{\mathrm{eff}}^{-1/2}$. Calibrating on the ordinary model gives

$$\widehat{D}_{\mathrm{drop}} = D_{\mathrm{ordinary}} \sqrt{\frac{n_{\mathrm{eff,ordinary}}}{n_{\mathrm{eff,drop}}}}. \tag{6}$$

Standard16's measured spread change predicts **21.2% smaller dips**. On the compact band-intervention panel, the minimum/long-mean ratio changes as follows:

Table 7: Minimum-to-mean margin ratio on the compact band-intervention panel.

	Ordinary	Dropout approximation	Dropout measured
Minimum / long-range mean	0.853	0.885	0.982

This retrospective approximation substitutes utility spread for amplitude spread. It explains part of the dip reduction, without predicting the increased mean. These ratios use the long-range mean; margin retention in Figure 6 uses the short-range mean.

L CONTEXT DEMAND, STOCHASTIC COMPARISONS, AND DISTANT INTERFERENCE

Cross-band coherence of the short-range comparison. The ordinary Standard4/8/16 sweep uses twelve paired seeds, width 64, two heads, two blocks and fixed frequency endpoints 1.7783×10^{-4}–1 rad/token. Increasing the count redistributes log-spaced nodes; the grids are not nested. The first-layer stationary positional Q/K numerator $K(d)$ retains signed phase, averages heads, and omits token terms, cross-band terms and input-dependent normalization. Figure 3 normalizes each seed by $K(1)$. Centered, unit-norm comparisons over lags 0–16 give paired cosines 0.971/0.987 for 4/8 and 8/16 bands, versus 0.132/$-$0.217 over lags 121–480. Ordinary median score half-height lags are 12.5/13/11; these are not attention half-widths. Thus agreement is local and gain-independent, and does not establish convergence to a common global kernel. All seeds are retained.

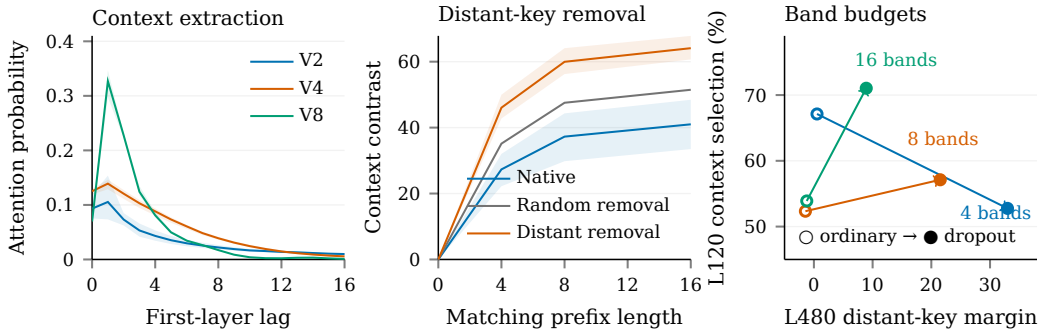


Figure 14: **Context construction, distant interference and band budgets.** Left: native first-layer source attention at L120 for three vocabularies. Middle: native matching-prefix response at L480, after removing distant content keys or a count-matched random subset. Shading: 95% seed-bootstrap intervals. Right: ordinary-to-dropout changes in local context-class selection and the lag-1-versus-distant score margin, twelve seeds per condition. Arrows connect cohort means, not training trajectories.

Alignment and the secondary context kernel. Predecessor alignment is central to induction circuits (Elhage et al., 2021; Olsson et al., 2022). We model its implementation through effective context filters in our harmonic models. The query for x_t contains x_{t-1}; the original source key contains x_t. For effective first-layer context weights a_q, a_k, prefix position $r \geq 1$ therefore corresponds to query lag $r-1$ and key lag r. With centered IID token features and fixed weights, a bilinear comparison has expected matching benefit proportional to $\sum_{r\leq m} a_q(r-1)a_k(r)$. A direct query-residual comparison adds a term proportional to $a_k(1)\mathbf{1}[m \geq 1]$. This makes lag 1 an informative alignment anchor and explains why a neighborhood can contribute. It does not uniquely specify a kernel: native routing, head mixing, residuals and normalization remain content dependent.

More generally, suppose consecutive outputs advance d source positions. Away from wrap boundaries, query lag ℓ contains $x_{t-d(\ell+1)}$, while a rival source key at $t+\Delta$, read at lag r, contains $x_{t+\Delta-r}$. Their tokens align when $r = d(\ell+1)+\Delta$. For fixed linear context filters, the geometric factor in their expected matching comparison is the channel-pair matrix

$$H^{(d)}_{hg}(\Delta) = \sum_{\ell\geq 0} a_{q,h}(\ell)\, a_{k,g}(d(\ell+1)+\Delta)\,. \tag{7}$$

Include a residual channel $a_0(\ell) = \mathbf{1}[\ell = 0]$ and both attention heads. Centered IID token covariance multiplies this geometric factor; the native readout, normalization and token-dependent routing need additional terms. Thus Equation 7 describes available matching relations, not a complete neural score. For ordinary copying $d = 1$, the residual-to-source route singles out lag 1. Multiple filters can supply additional compatible comparisons without uniquely determining their spectra.

Changing the required context alignment. We train 24 Sinusoidal16 models, four seeds each at V4/V8 and $d = 1, 2, 4$, for 8,192 updates. Source lengths are uniform over odd values 1–119;

output order is $x_{di \bmod L}$, so every source token is copied once. Source tokens, lengths, supervised-token counts and nonwrapping PE spans are matched across strides. Ordinary copying reaches 99.52/99.93% accuracy at V4/V8. The reordered tasks are harder: V4 stride 2/4 reaches only 35.55/34.18%, and V8 stride 2 ranges from 26.26% to 81.38%. Three of its four source profiles peak at 2. We retain these acquisition failures rather than treat missing peak movement as a clean test of alignment.

A separately registered continuation starts from all eight acquired ordinary models and adapts to every stride for 2,048 updates with inherited Adam state. All four V8 stride-2 models develop head peaks at *2 and 6*, but reach only 60.86% mean copying accuracy. These peaks implement two compatible routes: residual query to source lag 2, and query lag 2 to source lag $2(2+1) = 6$. Kernel movement is therefore not sufficient for complete copying. V4 reordered tasks remain poorly acquired, while V8 stride-4 outcomes vary.

We nominate token checks from native channel-pair overlaps before measuring fresh matched-lure outcomes. In 128 valid copied backgrounds per model, matching a wrong-payload rival at offsets 2 and 6 reduces the correct-minus-rival score margin by 6.47–7.58 and 5.65–7.76 units, respectively, across all four V8 stride-2 models. Matching both gives 12.93–15.32, whereas adjacent offsets 3 and 7 together give only -0.03 to $+0.02$. The per-background additive approximation has 6.1–9.7% relative RMS error. Three stride-4 models have only a substantial lag-4 check; a fourth also uses lag 16, consistent with its query-lag-3 route, with a 4.62-unit marginal effect. A nominal second overlap of negligible size is not counted as a useful check.

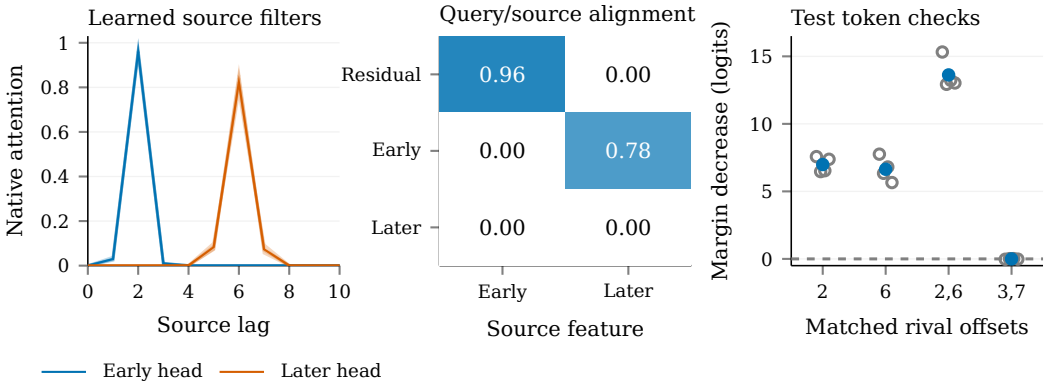


Figure 15: **Compatible context comparisons in stride-2 copying.** Left: source-attention profiles of all four V8 continuation seeds, with heads ordered by their source peak. Middle: their mean geometric alignment matrix, including the residual channel; entries are overlaps, not score amplitudes. Right: matching the nominated rival tokens changes native margins, whereas adjacent controls have negligible effects. Thin curves/open circles retain individual seeds. All models remain only partially acquired.

For k independent token checks, false candidates agree with probability V^{-k}. On 256 fresh stride-2 examples at a fixed interior query, a reference that guesses the modal payload among candidates agreeing on both checks gives 56.86% accuracy, versus 26.29% with one check; native models give 49.22–54.69%. This is an information-limited reference that ignores positional cues, not an upper bound or an exact accuracy prediction. The alignment account and the overlap nomination are distinct from a task-only acquisition law; the channel interpretation was developed after inspecting the learned profiles.

We measure $C(m) = \mathbb{E}[M_0 - M_m]$ by giving a wrong-token rival m matching prefix tokens, holding addresses, payloads and local query history fixed. The signed increment $C(r) - C(r-1)$ measures the additional contextual distinction, not a probability. Across 72 Standard4/8/16 ordinary/dropout endpoints, 12 seeds per cell, 8/16-band ordinary models reach 90% of their measured $C(40)$ at 6.60/7.21 tokens, versus 7.14/6.79 with dropout. Four-band dropout instead requires 12.77 tokens, versus 4.90 ordinarily. Normalized 8-versus-16 curve RMS differences are 0.049 ordinary and 0.018 dropout.

Transfer tests and stochastic demand. L120 prefix profiles predict mean responses to 64 scattered matching masks at an independent L240 geometry with normalized RMS errors 0.062/0.049 for ordinary/dropout Sinusoidal16, compared with 0.206/0.191 for uniform 64-lag weights. Target full/no-match anchors normalize gain: these are shape tests, not absolute score predictions. Softmax filters reconstructed from signed stationary coefficients give ordinary overlap-profile error 0.043 and dropout source-profile error 0.043. A universal product-of-filters rule is not established.

Weighted token-agreement scores also arise in soft induction constructions (D'Angelo et al., 2026); their match masks index agreeing context tokens, whereas our dropout masks index positional bands. We use a reduced margin model that additionally retains a matched-context residual and omitted interactions. Let Z_r be the random true-minus-rival agreement at lag r. The model is $M = P + \Gamma w_c^\top Z + E$, where P is a matched-context residual, Γ a gain and E omitted interactions. Their joint distribution over tokens and geometry matters; P is not a pure positional pathway. For a linearized comparison $M = \theta^\top F$,

$$\mathbb{E}[M] = \theta^\top \mu_F, \qquad \mathrm{Var}(M) = \theta^\top \Sigma_F \theta. \tag{8}$$

Covariance can change boundary-crossing risk at fixed mean separation. Under independent false matches of probability $1/V$, fully matched true contexts and nonnegative lag weights summing to one, the mean is $\Gamma(1 - 1/V)$ regardless of allocation. Mean-margin maximization alone cannot select the observed local profile. Available history, competitors, feature costs and their uncertainty must supply the nonexchangeable task structure.

For multiple rivals, retain the joint conditional law $\mathcal{L}(\mathbf{M}_c \mid L, V, \mathcal{G}, U, \theta)$, where $\mathcal{G}$ specifies candidate geometry, U the component mask and θ a trained state. The explanatory addressing risk is

$$\mathcal{R}(c) = \mathbb{E}_{X,\mathcal{T},U} \log \left(1 + \sum_j e^{-M_{c,j}(X,\mathcal{T},U)} \right), \tag{9}$$

where c parameterizes available comparisons and $\mathcal{T}$ includes length, vocabulary and rival geometry. This is not an identity with native token loss. Shared tokens and positional responses correlate rival margins. Addressing error is $\Pr(\min_j M_{c,j} \le 0)$ (counting ties as ambiguity), whereas softmax risk is Equation 9; neither is fixed by the mean margin or by marginal error probabilities alone. Component-dropout risk additionally averages over U. Variation over inputs and masks within a trained state is distinct from variation across trained seeds. We therefore retain paired candidate margins and condition on matching pattern and geometry, rather than impose one Gaussian residual law. Under an exchangeable IID-token null, a weighted sum of accidental matches has a discrete distribution even when its mean is allocation independent; its lower margin tail can still distinguish allocations. Native content-dependent routing need not follow that null distribution.

The joint support law also matters. Writing training risk as $\sum_U q(U)\mathbb{E}_X \ell(\theta; X, U)$ makes the full-support term explicit: equal per-band masking probabilities do not fix either $q(U)$ or its full-support probability. In the recorded Bernoulli schedules, uncorrupted updates comprise 33.8%/11.5%/1.8% for 4/8/16 bands; fixed-size quarter-band removal has none. Changing this law alters rehearsal of the native task as well as support tails, corruption covariance and RMS-gain fluctuations.

Resampling paired native baseline/gain anchors from 16 calibration backgrounds predicts mean confusion reasonably, but nominal 80% intervals cover only 58.2/56.0% of independent-background random-mask margins in ordinary/dropout models. Transferring each condition's empirical calibration distribution gives 70.7/72.5%. Adding centered singleton-response fluctuations raises coverage to 78.6/79.7% but worsens error-probability prediction and overcovers contiguous prefixes. Thus mean profiles and one unconditional noise law do not recover the full comparison distribution. Both heads remain separate in these checks.

Candidate coefficient selection. For a stationary score $K_c(r) = \sum_b c_b \phi_b(r)$, an effective source filter is $a_c = \mathrm{softmax}(\beta K_c)$. Source and overlap templates can be written $w_c = \mathrm{softmax}(\beta \Psi c)$, with $\Psi_b(r) = \phi_b(r)$ or $\phi_b(r-1) + \phi_b(r)$ respectively. If P, Γ, Z are held independent of c, the pairwise softplus risk obeys

$$\frac{\partial \mathcal{R}}{\partial c_b} = -\beta \, \mathrm{Cov}_{r \sim w_c} \left(\Psi_b(r), D_c(r) \right), \qquad D_c(r) = \mathbb{E}[\Gamma \sigma(-M_c) Z_r]. \tag{10}$$

This makes residual-error-weighted demand a candidate recruitment signal: ambiguous contexts contribute more than solved ones. A gain or parameter budget is needed to distinguish selection

from rescaling. The formula is not the native Adam update, which changes routing, gains, content features and values jointly. The task-length contrast in Equation 2 is a separately validated early selection bias, not a full prediction of this stochastic demand or of the exact learned kernel.

A chronological control probes ordinary Sinusoidal16 at 512/2,048/8,192 updates (four seeds each at V4/V8). Earlier derivatives with respect to physical band gain are compared with later changes in mean margin-deletion utility, using independent input panels. Address-risk demand has mean within-seed rank correlations 0.11/0.26 for the two V4 intervals and 0.29/-0.46 for V8; native token-loss demand gives -0.19/0.20 and 0.10/-0.09. Finite differences validate the local derivatives, but these do not reliably forecast accumulated recruitment. Here address risk is conditional on original source keys and averaged over both heads; it is not token loss. This control tests physical gain demand, not the stationary-coefficient derivative above. Native feature learning and optimizer dynamics remain necessary parts of a predictive acquisition model.

A finite-history stochastic selection model. One explicit way to break the exchangeable-lag null is limited matching output history. Let t be its available length, $p = 1/V$, $w_r \geq 0$, $\sum_r w_r = 1$, and $y_r = \gamma w_r$. Assume the true context matches at $r \leq t$, whereas beyond t both true and rival agreements are independent Bernoulli(p). The exact exponential margin-risk moment is then

$$\mathbb{E}e^{-M} = \mathbb{E}_t \left[\prod_{r \leq t} \left(p + (1-p)e^{-y_r}\right) \prod_{r > t} \left(1 - p + pe^{-y_r}\right)\left(1 - p + pe^{y_r}\right) \right]. \tag{11}$$

Unavailable history adds uncertainty rather than reliable separation. With token-exposure weights $\Pr(t) \propto 120 - t$, $t = 1, \ldots, 119$, minimizing the log of this moment over 64 weights gives localized solutions; making all lags available restores the uniform optimum. This is a convex reduced problem for fixed γ, not native token loss. It excludes first-token/EOS demands, positional residuals and adaptive marker features.

Retrospectively fitting one effective gain on four V4 seeds gives $\gamma = 5.19$; holding it fixed predicts 90% context-response lengths 12/8/6 for V2/V4/V8. On eight disjoint evaluation seed identifiers per vocabulary, the squared error between normalized cumulative profiles is 0.00712/0.00034/0.00054. Allowing vocabulary to change the solution improves V8 error from 0.00211 to 0.00054, but does not improve V2 over the fixed-V4 profile. Analytic gradients and the discrete moment were independently audited by finite differences and Monte Carlo. The effective gain is nevertheless fitted: inserting the measured V4 native score contrast 53.20 instead raises V4/V8 errors to 0.00881/0.00833. Thus finite history supplies a plausible localization constraint and a partial vocabulary prediction, not a calibrated native learning law or a derivation of unique Fourier coefficients.

Prospective history-exposure test. We freeze predictions before training 24 new Sinusoidal16 models: V2/V4/V8, four paired seeds, two supervision rules, and 8,192 updates. Both rules see identical IID inputs, lengths uniform over 33–120, and nonwrapping origins. Each example supervises exactly $L - 32$ content labels plus EOS: either a uniform subset of all copied positions, or only positions with at least 32 preceding output tokens. The architecture remains width 64, two heads and two blocks; Adam uses learning rate 10^{-3} and batch size 128.

Under Equation 11, all first 32 checks in the interior-supervision arm are exchangeable. Their contribution separates as $\sum_{r=1}^{32} f(w_r)$, where $f(w) = \log[p + (1-p)e^{-\gamma w}]$ is strictly convex for $\gamma > 0$. Consequently every optimum has $w_1 = \cdots = w_{32}$ and $\sum_{r=1}^{8} w_r \leq 1/4$, independently of gain calibration. The frozen gain predicts first-eight shares 0.747/0.880/0.972 for uniform supervision and 0.241/0.247/0.250 for interior supervision at V2/V4/V8.

At V8, all eight models acquire the common interior task. Nevertheless, the interior arm's measured first-eight share is 0.9946, not at most 0.25 (Table 8). Its paired change from uniform supervision is only -0.0020 $[-0.0030, -0.0010]$, versus a predicted -0.7220. The in-range probe gives the same localization. First-layer source attention does change: mean mass on lags 0–7 falls from 0.902 to 0.511. Each uniform model has local head peaks at 1 and 3; each interior model retains one strongly local lag-1 head, with a less local second head. Thus mean attention width and effective context-discrimination width are distinct.

This prospective result limits the identification of native comparisons with freely optimized nonnegative checks. Available history alone does not explain the learned effective profile; representational

Table 8: **Removing short-history supervision does not make the effective V8 context comparison broad.** All four paired seeds are retained. V2 interior accuracy is 68.75/68.16% for uniform/interior supervision; two V4 interior seeds fail acquisition. Their near-zero context gains make normalized ratios unstable, so Figure 16 retains raw contrasts for every seed.

	L120 interior accuracy (%)		V8 context share $C(8)/C(m_{\max})$	
Supervision	V4	V8	L120, $m_{\max} = 40$	L240, $m_{\max} = 64$
Uniform subset	99.39	99.91	0.9952	0.9966
Interior subset	69.99	99.85	0.9922	0.9946

constraints, positional interactions and optimization selection remain possible causes. It does not refute the alignment relation or distant-interference interventions. V2 and the failed V4 seeds cannot test an acquired optimum. All endpoints are fixed; we do not remove failed seeds or refit the gain. Four-seed intervals are descriptive and condition on the shared 32-background probe. The manipulation also changes supervision geometry and gradient assignment, so it does not isolate every task constraint.

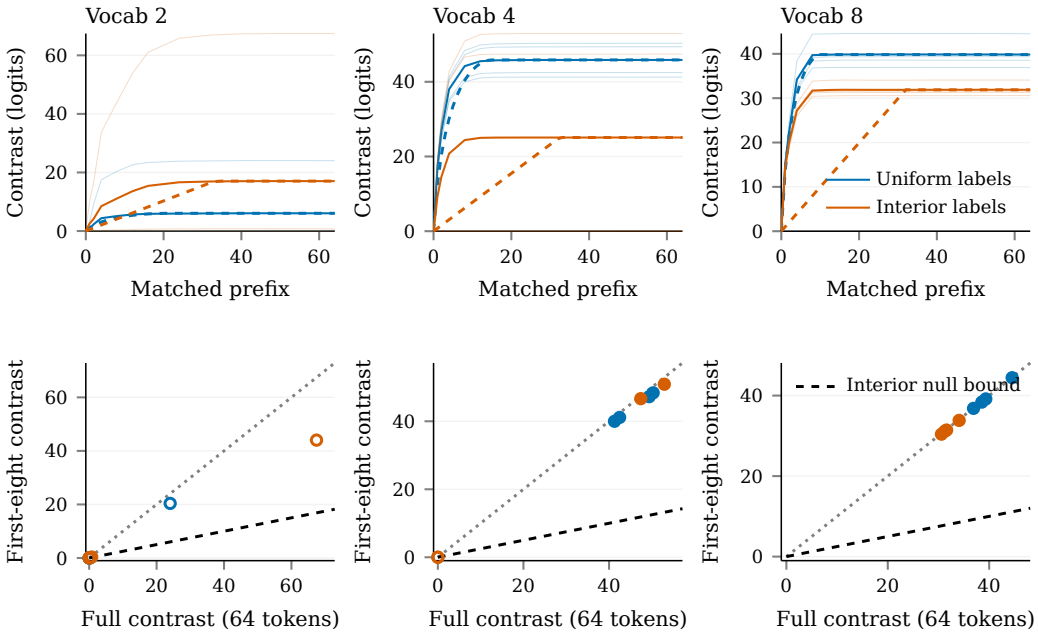


Figure 16: **Frozen demand predictions versus native context comparisons.** Top: primary L240 signed contrasts; thin curves show individual seeds, thick curves their means. Dashed curves are frozen predicted shapes, scaled to each arm's observed full contrast; absolute gains are not predicted. Bottom: every seed's first-eight versus full contrast, avoiding unstable ratios near zero. Filled/open symbols indicate at least/below 95% common-interior accuracy; no seed is excluded. The dashed line is the interior null's one-quarter bound under the proposed additive identification, and the dotted line is equality.

Native stochastic loss demand and independent finite edits. We probe 48 Sinusoidal16 models at updates 8,192 and 65,536: ordinary $V = 2, 4, 8$ and V4 dropout, twelve seeds per cell. Actual variable-length copying loss, including EOS, supplies the score gradient. For first-layer normalized scores s_{hij}, define signed lag demand $\rho_h(d) = -\sum_{i-j=d}(\partial \ell/\partial s_{hij})/(\sigma_i \sigma_j)$, where σ_i, σ_j are the input LayerNorm standard deviations. For a virtual stationary coefficient with harmonic feature ϕ_b, the negative gradient is its Fourier projection $\sum_d \rho_h(d)\phi_b(d)$, multiplied by the squared mask gain when that component is corrupted. This is an exact chain-rule measurement conditional on the native network, not a task-only learning law. Signed loss demand is neither an attention density nor a desired kernel, and its instantaneous peak need not be lag 1.

Four batches of 32 examples select a correction direction. We freeze it and evaluate a small stationary-score edit on 64 new batches per condition (2,048 sequences and 64 new mask draws), at the final checkpoint. The primary edit has centered stationary-numerator RMS 0.01 over lags 0–242; this does not fix the change in every normalized native score. No parameters are trained. The larger confirmation gives:

Table 9: Held-out stationary-score edits at the final checkpoint. Cross-entropy (CE) is measured in nats; changes compare the edited and unedited masked model.

Training	Clean CE	Masked CE	Masked ΔCE	Improved seeds
V2 ordinary	0.09938	6.20318	$+0.01274$	8/12
V4 ordinary	0.00101	1.11448	-0.01105	12/12
V8 ordinary	0.00058	0.46662	-0.00456	12/12
V4 dropout	0.00283	0.00434	-0.00003	3/12

The V4/V8 masked changes have paired seed-bootstrap 95% intervals $[-0.01360, -0.00850]$ and $[-0.00569, -0.00347]$ nats; reversing the edits increases loss by 0.01357/0.00493. Removing the direction along the current centered kernel still reduces masked loss by 0.00941/0.00414, so a uniform increase in that kernel's gain is not the only available correction. V2's mean change and the dropout masked change have intervals including zero. Intervals condition on shared evaluation inputs and masks. The initial four-batch calibration does not reliably predict change magnitudes.

Ordinary clean acquisition can therefore leave substantial masked comparison demand, with reproducible directions for reducing it. Dropout-trained states have much less remaining mask-induced loss. These score-path edits do not establish that native optimizer updates follow the same direction, or that reducing this demand mediates extrapolation. The Fourier-gradient audit has maximum relative discrepancy 7.1×10^{-5}. An initial float32 confirmation failed its full-logit parity gate at unused padding positions; its results are preserved. Repeating every model and condition in float64, with identical frozen directions and inputs, gives native/zero-edit discrepancy below 4.1×10^{-11}.

Vocabulary validation. Unrelated IID prefixes agree with probability V^{-m}. A union bound over N rivals gives at most NV^{-m}, without requiring independent candidate events. Fixing this ambiguity suggests a scale proportional to $1/\log V$ and the descriptive frequency dilation

$$F_V(\omega) \propto F_4(\omega/a), \qquad a = \log V / \log 4. \tag{12}$$

Four V4 reference seeds and eight disjoint evaluation seed identifiers per target vocabulary give broad-spectrum error reductions of 32.5% for V2 and 48.7% for V8. The V2 all-seed improvement interval includes zero; V8's is positive. These are signed context-margin deletion utilities, smoothed by one log-frequency spacing; the individual-band prediction is weaker. Native cumulative-response 90% crossings are eight/eight/five tokens in acquired V2/V4/V8 cohorts. No single dilation explains every cumulative and incremental response. The reference/test split is retrospective.

The new routing measurement includes all 12 ordinary seeds at each vocabulary, including acquisition failures. At the true source key in L120, mean lag-1 attention is 0.105/0.139/0.327 for V2/V4/V8; mass on lags 0–7 is 0.459/0.758/0.940. Cohort means peak at 1, but V2 individual peaks often occur at 0. These are native content-dependent attention profiles, not stationary positional-score coefficients or physical-band deletion amplitudes.

Distant interference and an exact attention identity. For a first-layer query, partition causal keys into removed distant keys $\mathcal{F}$ and retained keys $\mathcal{R}$. With native scores s_j,

$$G = \log \sum_{j \in \mathcal{R}} e^{s_j} - \log \sum_{j \in \mathcal{F}} e^{s_j}, \quad \epsilon = \sigma(-G), \quad h = (1-\epsilon) h_{\mathcal{R}} + \epsilon h_{\mathcal{F}}. \tag{13}$$

Removing $\mathcal{F}$ and renormalizing gives $h_{\mathcal{R}}$ exactly before output projection. More distant keys increase their log-partition even without one dominant alias. With independent distractor content, useful query/key context covariance is attenuated by their retained masses; shared tokens and adaptive routing instead require their joint distribution. This is a mechanism for damaging context comparison despite intact local positional distinctions.

Boundary markers can remain useful at a distance. Separating their mass m from content gives the more specific identity

$$h = mh_{\text{markers}} + (1-m)\big[(1-\epsilon_C)h_{\text{near content}} + \epsilon_C h_{\text{far content}}\big], \tag{14}$$

where ϵ_C is distant mass conditional on content keys. Content-only removal at fixed marker weights tests interference without erasing landmarks.

Causal protocol and outcomes. The new probe evaluates 48 Sinusoidal16 endpoints: ordinary V2/V4/V8 and V4 dropout, 12 seeds each, at L120/240/480. True and rival indices are $\lfloor L/3 \rfloor$ and $\lfloor 11L/15 \rfloor$, with different payloads. The rival has 0/4/8/16 matching prefix tokens. Thirty-two backgrounds at two nonwrapping origins are shared within vocabulary. Local243 keeps causal lags 0–242 and BOS/separator. Random243 keeps the same number of keys in every row, retaining the nearest 32 keys and markers. Only first-layer access changes. Every L120 output remains exactly native, and all restoration checks pass.

For 16 matching tokens, $\mathcal{G}$ contains every source address with the same 16-token prefix as the query. True and planted-rival addresses are included; accidental matches remain. The exact factorization $P(\text{true} \mid \text{source}) = P(\mathcal{G} \mid \text{source})P(\text{true} \mid \mathcal{G})$ holds per example/head, not as a product of cohort averages.

Table 10: Context-class selection, within-class selection and context gain at L480 under first-layer attention interventions.

Vocabulary / training	Intervention	Class selection	Within class	Context gain
V2 ordinary	Native	0.87%	63.70%	29.95
	Local243	1.44%	57.80%	39.46
V4 ordinary	Native	32.30%	49.82%	41.00
	Random243	38.26%	53.46%	51.48
	Local243	45.96%	44.58%	64.09
V8 ordinary	Native	91.57%	42.54%	60.57
	Local243	99.75%	64.36%	87.64
V4 dropout	Native	65.38%	59.83%	61.58
	Local243	65.38%	59.83%	61.59

At L480, V4 Local243 increases class selection by 13.66 points (paired 95% interval [6.24,21.70]), or 7.70 points [3.16,12.36] over Random243. Context advantage increases 23.10 [15.27,31.06] score units. V8 class selection rises 8.17 points [1.90,17.06]; V2's 0.58-point gain is uncertain. On ambiguous V4 examples, queried-token accuracy falls from 47.66% to 40.36% despite improved context selection. Context repair does not solve occurrence discrimination. All heads/seeds/vocabularies remain in the analysis; seed intervals condition on shared evaluation backgrounds.

A separate IID-copy cohort establishes behavioral relevance: ordinary L480 teacher accuracy rises 65.00→93.74%, versus 78.47% with Random243; stopped generation rises 30.73→50.27%, versus 38.34%. Exact content plus EOS rises from 0 to 6 of 384 generations. Dropout endpoints are nearly unchanged by either repair. Autonomous repair is partial. Distant positional-score donation gives 90.58% teacher accuracy for cross-frequency terms, 93.15% for all positional terms, and 66.52% for stationary terms alone. These nonlinear effects are not additive mediation shares. Native scores also contain nonstationary and mixed content/position terms.

Native adaptation and controlled attention-mass transfer. We fork each of twelve ordinary V4 Sinusoidal16 endpoints at update 65,536 into six continuations: clean or 25% component-masked training, updating all parameters, only the first block's Q/K projection weights and biases, or only the second block and final normalization/readout. Every arm uses the same 512 fresh batches of 128 examples, original Adam moments and learning rate 0.001. The downstream-only arm freezes embeddings and the whole first block; the Q/K-only arm freezes every other parameter, including V projection rows. These are different adaptation locations, not equal-capacity controls.

Independent short evaluation uses 512 sequences per clean/masked condition and sixteen new component masks. IID content-token accuracy uses 64 examples at each of L120/480. Masked training

reduces held-out masked CE from 1.064 to 0.037 with all parameters, 0.954 to 0.228 with Q/K only, and 1.097 to 0.257 downstream only. Respective paired L480 gains over clean continuations are 21.51 [12.58,30.70], 16.30 [8.95,23.66], and 2.86 [1.28,4.37] percentage points. Short accuracy decreases by 0.16/0.33/0.32 points; these costs and every seed remain in the analysis. Downstream-only training leaves all measured first-layer scores and routing exactly unchanged. Thus mask robustness can improve without learning a new first-layer kernel, but this route gives much smaller long-copying gains in the tested budget.

We next transfer only near–far *content* competition between paired endpoints. Each donor computes native first-layer scores on its own features; each recipient retains its embeddings, Q/K/V, normalization, within-group attention and downstream computation. We swap the near/far content logpartition gap at lag 243, with a common content-score correction preserving *every marker attention weight*. Equation 16 gives the resulting mixture. No feature vectors or parameter coordinates are exchanged. The following 108 paired tests include all three training scopes and swaps on source, output-query, or both row sets.

Table 11: L480 copying accuracy after near–far content-competition transfer across adaptation scopes.

Adapted parameters	Clean	Masked	Clean + learned gap	Masked + old gap
All	64.55%	86.08%	82.92%	70.12%
First-layer Q/K	65.57%	81.73%	81.05%	66.74%
Downstream only	65.18%	68.00%	65.18%	68.00%

On 64 fresh IID L480 inputs, full-network adaptation gains 21.53 [12.50,30.73] points. Forward competition donation gains 18.37 [10.69,26.21]; restoring old competition costs 15.96 [8.55,23.79]. Source-only and query-only forward gains are 2.40 [0.85,4.24] and 12.61 [6.64,18.88], with a positive joint interaction of 3.37 [1.48,5.52] points. Reverse restoration costs 3.96/15.02 points by role and has interaction -3.02 [-5.12,-1.28]. The interaction is a finite difference of controlled interventions, not a unique mediation percentage. Unlike the eight-band acquired-model repairs, source-side changes make a detectable contribution here.

On 32 fresh paired prefix backgrounds, full-network forward donation improves matching-class selection by 9.63 [4.19,15.82] points, with 7.87 [3.42,12.95] from query-only donation. The Q/K-only scope gives 15.48 [8.99,22.11] points of forward copying gain and 7.26 [2.46,13.07] points of class-selection gain. Downstream-only transfers are exact identities. All swaps preserve their recipient's L120 outputs. Independent direct probability rescaling agrees with the logit implementation within 1.3×10^{-12} on fresh short/long inputs, and query-only swaps leave all second-layer source Q/K/V exactly fixed. Thus the pathway extends to native dropout training with a learned value and feature system; it remains an endpoint causal diagnostic using test-length information, rather than a short-only predictor or unique optimization mediator.

The earlier Q/K-only experiment also transplanted score entries: learned far scores alone lower accuracy from 65.53% to 56.50%, whereas learned local scores alone raise it to 84.66%. Thus the useful change is *relative* local–far competition, not necessarily absolute distant-score suppression. First-layer score changes include content and nonstationary positional terms.

In an earlier all-key Q/K-only experiment, including distant markers, we further permute target mass across the eight examples in each evaluation batch, preserving its empirical marginal distribution at each head/query row but breaking assignment to the current tokens/origin. Accuracy remains 82.41%, a 0.03-point change [-0.04,0.12]. Replacing targets by their batch mean at each head/query row gives 82.69%, a 0.31-point gain [0.11,0.54]. Thus exact input-specific mass assignment is unnecessary in this panel; head/row-conditioned means retain the effect. These are oracle diagnostics at the evaluation length, not a deployable short-to-long predictor or a single global mean. Within-seed variance across backgrounds at fixed head/query row accounts for 6.45%/4.93% of summed clean/learned mass variance; most variance is across head/row conditional means. This does not make downstream rival margins deterministic. All mass interventions preserve short outputs exactly; maximum float64 native-output reconstruction error is 3.2×10^{-12} and mass-matching error is below 1.9×10^{-15}. For the earlier all-key full-model transfer panel, native-output reconstruction error is below 4.0×10^{-12} and mass error below 1.3×10^{-15}. An independent audit of all twelve

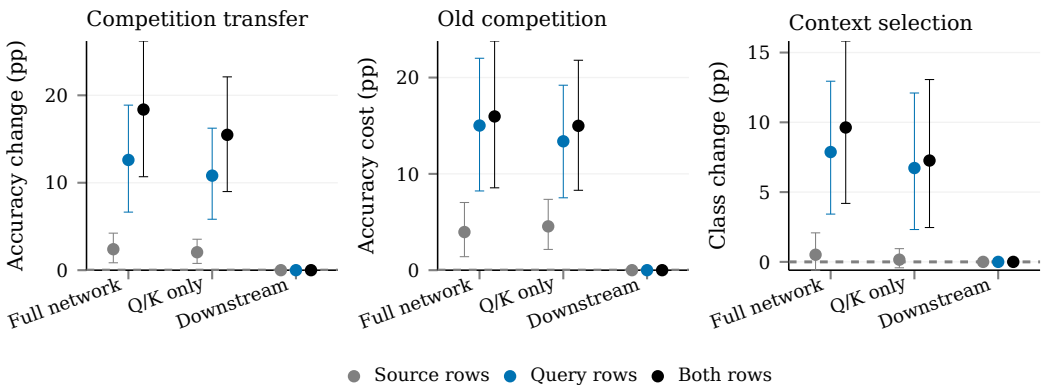


Figure 17: **Native dropout changes context construction through distant content competition.** Left: learned competition donated to the clean recipient. Middle: accuracy cost of restoring old competition in the masked recipient. Right: matching-class selection under forward donation. Every marker weight and recipient value remains fixed. Source, query and joint row interventions retain all twelve paired seeds in each scope; intervals are 95% seed-bootstrap intervals conditional on the evaluation panels. Downstream-only controls are exact identities.

full-model pairs verifies unchanged recipient parameters, unchanged within-group conditional attention, and the exact pre-projection mixture identity. The donor supplies mass at the test length: this intervention does not identify a task-only learning law or show that the mechanism is exclusive to component dropout.

Dropout under finite band budgets. Extending the native probe to Standard4/8 gives 72 V4 endpoints, 12 seeds per band-count/recipe cell. We measure $g_1 = \log A(q, q-1) - \log \sum_{q-j\geq 243} A(q,j)$ before averaging. This native diagnostic includes distant marker keys; the causal exclusion above retains them. It is not the exact margin of the removed set.

Table 12: Paired dropout-minus-ordinary changes across band budgets. Brackets give 95% seed intervals; accuracy changes are in percentage points (pp).

Bands	Δg_1	Δ L120 class (pp)	Δ L480 teacher (pp)
4	+32.44 [19.94,44.13]	-14.36 [$-26.67, -3.87$]	-6.05 [$-26.72, 14.58$]
8	+22.97 [21.66,24.07]	+4.78 [$-4.48, 15.68$]	+41.01 [34.58,47.02]
16	+10.13 [9.58,10.66]	+17.14 [6.21,28.54]	+33.92 [19.95,47.42]

Entries are paired dropout-minus-ordinary changes with 95% seed intervals. Teacher accuracy averages copied tokens across the four planted-prefix conditions, not IID generation. Raw short context gain falls by 8.90/9.99/4.32 score units for 4/8/16 bands, but raw gain also depends on scale. The significant four-band class-selection cost supports a restricted-support tradeoff; richer supports can improve both local selection and distant suppression. A universal claim that dropout must sacrifice context capability is unsupported.

Band budget and boundary references. Separate mass and conditional-routing transfers test all 36 ordinary/dropout pairs. Donating learned far mass improves ordinary 8/16-band IID L480 accuracy by 30.44/30.25 points, but changes four-band accuracy by -8.31 [-17.29,0.01]. We therefore ablate distant content and markers within each native model. All 72 endpoints, 64 new IID examples per length, and both heads remain included; every short output is unchanged.

At fixed marker weights, content removal gains 0.20 [-0.87,1.35], 24.76 [17.92,31.64], and 29.57 [17.30,41.96] points. Removing distant markers instead costs four-band models 13.24 [5.96,20.99] points. Their native distant attention contains 9.50% content and 16.14% markers, versus 29.82% and 2.02% with sixteen bands. Dropout endpoints are nearly unchanged by content removal. Thus total distant mass is not a universal error predictor: boundary references and content interference

Table 13: L480 copying accuracy of ordinary models under distant-content and distant-marker ablations.

Ordinary bands	Native L480	No far content, fixed markers	No far markers
4	70.30%	70.50%	57.06%
8	36.04%	60.80%	36.76%
16	64.84%	94.41%	65.42%

must be distinguished. Transplanting conditional local attention damages short class selection in both directions at every budget; it does not isolate the four-band context cost from coadapted values/readout.

A realizable native comparison repair. Virtual coefficient edits need not be realizable by native Q/K weights. For the additive codes here, let R span the centered token and positional features after multiplication by the first LayerNorm gain. Its normalized input has the exact form

$$x = \beta + Ry, \qquad \phi = (y, 1), \qquad s_h(q,k) = \phi_q^\top F_h \phi_k, \quad F_h = A_h^\top B_h / \sqrt{32}. \tag{15}$$

The input-dependent normalization denominator changes y, but not this span. With seven token embeddings and K repeated sine/cosine pairs, centering gives $\dim \phi \le 7 + (2K-1) + 1$. The measured dimensions are 15 and 23 for four/eight bands. Every metric on either span therefore factors into the existing 32-dimensional Q/K vectors per head. Given an SVD of F_h, square-root singular values give A_h, B_h; mapping their linear parts through $R^\top$ and adjusting the biases realizes the comparison in native weights. Values, embeddings, normalization and downstream parameters remain fixed. This unrestricted factorization is not asserted for sixteen bands, whose span can exceed 32.

We fit short native attention and a long target that reduces distant-content weights by 90%, redistributing mass to near content while preserving marker weights. Calibration uses 32 short and 16 long examples, with no token-label objective. Both predefined short/long weights, 10 and 100, remain included. Each head's metric displacement is clipped to 25% of its native Frobenius norm, excluding the query-only score gauge. The fitted native metric only approximates the marker-preserving target. This is a feasible construction, not a certified optimum or a short-only learning recipe: it uses long geometry.

Across twelve eight-band seeds, long accuracy improves by 8.35 [5.44,11.63] and 12.07 [7.75,16.96] points at weights 10 and 100. The latter raises accuracy from 35.97% to 48.04%, improving every seed; short token accuracy changes by 0.010 [-0.001,0.025] points and short matching-class selection by -0.148 [-0.363,0.049]. Long matching-class selection rises by 10.07 [5.16,15.84] points. Four-band mean token gains are approximately zero, consistent with their weaker distant-content ablation response. Neither their failed repair nor the numerical optimizer establishes a necessary tradeoff. Eight-band masked validation loss increases by 0.122 [0.085,0.156] nats/token: improved clean extrapolation does not imply improved component-corruption robustness in this intervention. Independent checks cover all 72 native checkpoints, unchanged non-Q/K parameters, and the full supplied feature span. The zero-edit refactor preserves measured token accuracy exactly.

Short-only preservation under component removal. The same native metric construction permits a short-only intervention: preserve the parent's clean attention distribution when PE components are removed. We fit 32 variable-length inputs of lengths 1–120 under 16 nonempty Bernoulli-25% support draws, with RMS rescaling. Both predefined clean-fit weights 10/100 and all twelve parents per four/eight-band basis are retained. The objective is clean-attention KL times this weight plus masked-attention KL to the clean teacher, under the preceding displacement budget and solver. Short-only fits are frozen before constructing any long fitting panel.

On independent inputs and support draws, eight-band masked attention KL falls from 0.0896 to 0.0825/0.0806. L480 token accuracy rises from 35.97% to 38.29/39.07%, paired gains 2.32 [0.89,3.71] and 3.11 [1.07,5.14] points; short changes are -0.001/$+0.003$ points. Masked token loss also falls by 0.189/0.319 nats. Four-band masked KL and token loss improve, but long accuracy changes only 0.19 [-0.16,0.72]/0.02 [-0.07,0.13] points. Thus short-support robustness can improve extrapolation in the native model, but does not determine it across bases. Combining this

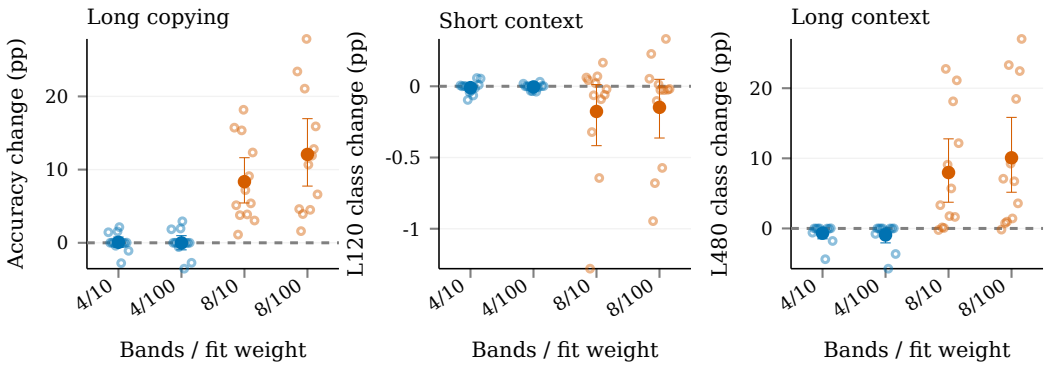


Figure 18: **The interference mechanism guides a native comparison repair.** Changes from ordinary parents after modifying only first-layer Q/K weights. Open circles retain all twelve seeds per basis; filled circles and intervals show paired means and 95% seed-bootstrap intervals. Both predefined fit weights are shown. Long geometry guides fitting; held-out examples evaluate copying and context selection. Width, heads and positional bases are unchanged.

objective with the long-geometry target gives eight-band gains of 1.66/5.48 points, smaller than the long-only repair; finite clipped fits do not establish a necessary tradeoff. These are acquired-model interventions, not a derivation of initial recruitment or a complete mediation analysis of dropout training.

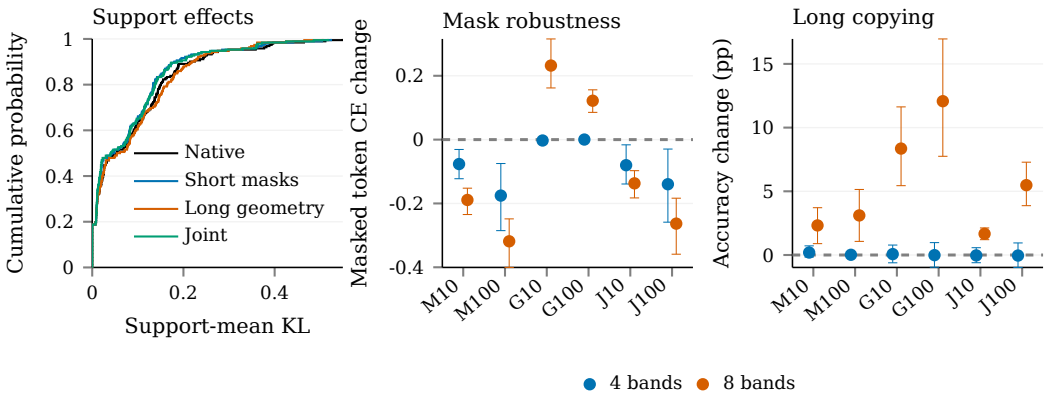


Figure 19: **Support robustness and long-range copying are related but distinct.** Left: eight-band distribution of support-mean attention KL to the clean teacher, with clean-fit weight 100; equal weight for each seed, keeping all 16 independent mask draws. Middle/right: paired native outcome changes for both fixed weights, with 95% seed-bootstrap intervals conditional on evaluation panels. M: short masked-attention preservation; G: long-geometry oracle; J: both. Numbers denote clean-fit weights. Only M uses exclusively short fitting inputs. All twelve seeds per basis and all settings are included.

Distant competition and output-query construction. Partition first-layer keys into markers, near content and content at lag ≥ 243. With marker mass m, far fraction r of content attention, and conditional value averages h_M, h_N, h_F, the head output before its output projection is exactly

$$h = mh_M + (1-m)\big[(1-r)h_N + rh_F\big], \qquad r = \sigma(\log Z_F - \log Z_N). \tag{16}$$

Here Z_N, Z_F are the near/far content score partitions. Fixing marker weights and within-group attention gives $\Delta h = (1-m)\Delta r(h_F - h_N)$. Its expectation depends on the joint distribution of the mass change and conditional features, not mean r alone.

For all 144 parent–repair pairs above, we transfer this per-input/head/row competition in both directions, preserving every marker weight, each conditional content distribution, and recipient values.

On 64 fresh L480 inputs, eight-band short-only repairs gain 2.44/3.21 points. Transferring only competition to the parent gains 1.18 [0.12,2.20]/1.56 [0.07,3.01] points; restoring old competition costs 1.60 [0.38,2.80]/2.11 [0.36,3.83]. Long-geometry repairs gain 8.60/12.33 points, with forward competition gains 7.61/11.20. Four-band transfer effects remain near zero. These contributions need not add to the native gain and are not unique mediated fractions.

Another 288 paired evaluations restrict the swap to source or output-query rows. Eight-band query-only gains are 1.19/1.57 points for short-only repairs and 7.64/11.24 for long-geometry repairs; source-only effects are within 0.02 points of zero. Independent activation checks show query-only swaps leave *all second-layer source Q/K/V exactly unchanged*. Thus their effect on source-restricted comparisons changes the query with fixed source keys, connecting this failure mode to query rescue. Token outputs can also use changed output-side keys/values, so this is not a pure Q-only token test.

The role difference follows unequal causal exposure: ordinary eight-band far-content fractions average 0.0082 on source rows and 0.5647 on output-query rows. The query predicting source token t is at row $L + t$, and therefore encounters more distant preferences. On the fresh matched-context panel, query-only long-geometry swaps improve class selection by 6.73/8.50 points; short-only class effects remain uncertain. Its correct source row 161 has no lag-243 keys, so role dominance is conditional on this geometry. All swaps are exact identities at L120. The donor is evaluated at test length; the swaps are causal diagnostics, not short-only prediction methods.

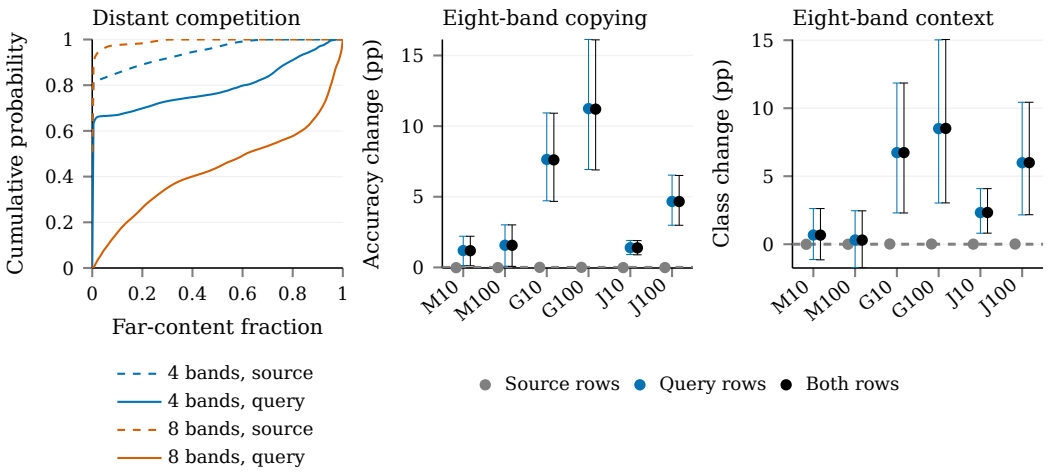


Figure 20: **Distant interference primarily affects output-query construction in these eight-band repairs.** Left: equal-seed distributions of native per-input/head/row far-content fractions at L480. Middle/right: forward competition swaps restricted to source, query, or both roles; 95% paired-seed intervals, twelve seeds. M/G/J and fit weights are as in Figure 19. Source-only effects are near zero; query-only effects closely reproduce the full competition swap.

Corruption onset, optimizer state, and support law. We additionally run 96 paired four/eight-band continuations from acquired ordinary endpoints, matching the existing sixteen-band forks above. With inherited Adam at 10^{-3}, four-band masked adaptation reduces short accuracy from 99.88% to 49.12%. This collapse cannot establish an unavoidable context–margin tradeoff. Across all twelve four-band parents, reducing the rate to 10^{-4} gives 97.30% short accuracy; resetting Adam at 10^{-3} gives 94.48%, with unchanged starting weights, masks and training inputs. Neither restores the paired long-copying performance in this 512-update budget.

At the same nominal 25% per-band rate, removing exactly a quarter of the bands instead of a Bernoulli subset improves eight-band long accuracy by 4.13 [2.01,6.41] points, but gives no clear sixteen-band benefit. It does not prevent four-band collapse at 10^{-3}; at the stabilized 10^{-4} rate it instead reduces short/long accuracy by 4.74/3.24 points relative to Bernoulli masking. All twelve seeds and predefined conditions are retained. Thus native adaptation depends on optimizer state and the joint support distribution, including clean-task rehearsal; marginal masking rate and band count

alone do not specify it. These controls do not isolate a single higher moment or establish a universal preferred masking law.

A further 24 continuations preserve every clean-rehearsal step and each band's total survival count, while replacing 23 singleton/triple-support pairs by two size-two supports. Removing singleton supports still does not prevent collapse at 10^{-3}: short accuracy is 48.08%, versus 49.12% originally. At 10^{-4}, long accuracy improves by 1.39 [0.10,2.97] points, with short change -0.07 [-0.32,0.17]. Thus the harshest supports are not a sufficient explanation of collapse in this protocol. Co-survival, RMS gains and within-pair timing still change; the control does not isolate a unique support statistic.

Local moments do not predict finite adaptation steps. At each of the 36 acquired parents, we sample 32 independent minibatch/support pairs under clean, Bernoulli and fixed-size corruption, with inherited or reset Adam. Every trial restarts from its parent. For a measured update δ with mean μ and covariance Σ, the local approximation is $\mathbb{E}\Delta\mathcal{L}_0 \approx g^\top\mu + \frac{1}{2}\mu^\top H\mu + \frac{1}{2}\operatorname{tr}(H\Sigma)$. Derivatives use 32 calibration examples and outcomes 64 independent examples. For four-band Bernoulli corruption with inherited Adam, the observed one-step clean-loss increase has median 0.070 and 90th percentile 1.475 nats/token. Its mean is 0.419, versus a quadratic forecast of 0.047, despite covariance contributing 82.3% of the quadratic term. Thus retaining stochastic moments matters but is insufficient at the native step size. This diagnostic conditions on the actual update; it neither predicts the update itself nor establishes a multi-step recruitment law. On a separate pair of 32-example panels, retaining the exact output loss of linearly extrapolated logits $z + J\delta$ reduces full-step relative RMS error from 0.95/0.97/0.92 to 0.75/0.81/0.77 for inherited-Adam Bernoulli updates at 4/8/16 bands. All twelve seeds and eight predefined draws per condition remain included. Substantial error remains even with the output nonlinearity intact; nonlinear feature changes also limit this local approximation.

Our account is that support corruption exposes comparisons ordinary acquisition can leave fragile. With few bands, local discrimination and distant separation can conflict; adequate supports can satisfy both more reliably. These cohort results do not establish a Pareto-optimal solution or isolate optimization mediation. Cross-frequency score masking's early teacher benefit did not survive its fixed 65,536-update endpoint: L480 teacher gain over ordinary was 2.96 points $[-3.01, 9.63]$, and stopped generation decreased by 3.17 points $[-6.03, -0.51]$. Targeting one positional interaction alone is not a demonstrated replacement for full component dropout.

M EXPANDED CONNECTIONS TO PRIOR WORK

Sinusoidal positional encodings supply fixed harmonic features (Vaswani et al., 2017), while RoPE introduces relative-position structure through rotations of queries and keys (Su et al., 2024). Learnable Fourier maps parameterize spatial positional encodings (Li et al., 2021), and FourierLearner-Transformers learn spectral representations of relative positional mechanisms (Choromanski et al., 2024). Fourier features also approximate stationary kernels (Rahimi & Recht, 2007) and alter the effective bandwidth of coordinate-based MLPs (Tancik et al., 2020). Spectral-bias studies analyze frequency-dependent learning (Rahaman et al., 2019); modular-addition transformers learn task-structured Fourier circuits (Nanda et al., 2023). We study which positional frequencies and comparisons a task makes useful, including compensation among different acquired mixtures.

Analyses of trained RoPE models identify different uses of high and low frequencies (Barbero et al., 2025), bands shaped by base and training length (Oka et al., 2026b), and per-dimension utilization (Oka et al., 2026a). FoPE studies frequency-domain effects on length generalization (Hua et al., 2025); Resonance RoPE addresses positional distribution shift through modified rotary features (Wang et al., 2024). Closest to our acquisition question, Wu et al. (2026) relate RoPE frequency usage to the distance structure of training data. Our additive-code experiments derive a spectral contrast from copying geometry and test how length and vocabulary shape learned comparisons, including local agreement and divergent distant responses.

Induction circuits match preceding context and copy the associated continuation (Elhage et al., 2021; Olsson et al., 2022). Jelassi et al. (2024) give a two-layer copying construction whose error is controlled by repeated n-grams. Generalized induction theory analyzes learning informative context features under a staged training procedure (Chen et al., 2024); D'Angelo et al. (2026) characterize soft comparisons over partial matches. Song et al. (2025) study compatible representations between earlier and later copying circuits. Repeated-context ambiguity is also explicit in Wen et al. (2026). These mechanisms specify the demands in our controlled task. We connect those demands to learned harmonic comparisons and use planted ambiguities to measure which distinctions the acquired bands support.

Task-dependent algorithm selection also has precedents: gradient-based learning can recover informative dependencies (Nichani et al., 2024), and trigger-distance diversity can favor induction over positional shortcuts (Kawata et al., 2025). Length-generalization studies compare positional schemes (Kazemnejad et al., 2023), relate generalization to short RASP-L programs through a conjecture (Zhou et al., 2024), and analyze identifiability under idealized norm-regularized inference (Huang et al., 2025). Our controlled comparisons isolate learned use of a harmonic basis and its task-dependent constraints within a fixed architecture.

Randomized positional encodings improve length generalization by exposing models to sampled positions from longer sequences (Ruoss et al., 2023). Our main interventions instead hold the frequency basis and positional-span protocol fixed while changing band availability during training. Dropout trains thinned networks (Srivastava et al., 2014), with an adaptive-regularization interpretation in generalized linear models (Wager et al., 2013). DropPos removes patch positional embeddings and learns to reconstruct positions (Wang et al., 2023). Our corruption removes complete harmonic pairs, restores positional RMS, and retains the copying objective. We test how this changes learned band alternatives; the conditional subset-margin theorem specifies when subset competence and complementary separation protect long-range margins.

N FROZEN FIRST-LAYER FORECASTS AND BEHAVIORAL INTERVENTIONS

The experiments separate three endpoints: predicting a trained first-layer score, associating that prediction with copying across models, and intervening on forecast-selected attention to change an actual output. All use the same 216 existing endpoints from the six-grid study (six bases, three dropout rates, twelve seeds); they are additional measurements, not independent training replications. Checkpoint hashes match the original behavioral evaluations. All architectures remain width 64, two heads and two blocks. There is no acquisition filtering, retraining or fitting to long-copying labels.

N.1 WHAT IS KNOWN BEFORE A DISTANT-SCORE FORECAST

For each first-layer head, write $M(q, d) = S(q, q-1) - S(q, q-d)$, so positive margin favors the predecessor over a candidate at lag d. The native independent-content target averages scores exactly over all sixteen pairs of four content tokens, retaining embeddings, affine biases and input-dependent normalization. Separate targets are the positional numerator and the native average over four matching-token pairs. These averages of logits are not expected attention probabilities or copying error probabilities.

The conditional signed predictor fits

$$\widehat{M}(q, d) = \sum_b \{c_{qb}[\cos \omega_b - \cos(\omega_b d)] + s_{qb}[\sin \omega_b - \sin(\omega_b d)]\} . \tag{17}$$

Each test query supplies its own short-lag score row; projection weights are not supplied to the regressor. Of 64 absolute query coordinates from 481 to 1023, alternating sets of 32 choose ridge regularization and evaluate the forecast. On calibration queries, lags 1–160 fit and 161–240 select among 13 ridge strengths from 10^{-12} to 1. Final fits use lags 1–240; the primary test uses 243–480. The first-layer representation depends on the current token and position, so short probes can measure these rows without a long copying execution. Random origins expose these absolute coordinates during training: this extrapolates lags, not unseen absolute coordinates.

The original 1–120 calibration experiment remains archived. After inspecting it, the follow-up used the full training-supported content span, 1–240, while keeping the distant test fixed. Every follow-up prediction was then frozen before its distant inner products were evaluated. This sequencing prevents long-score parameter fitting; it does not make the follow-up an independent prospective replication. Analytic features evaluated at 80 and 160 decimal digits agree after conversion to float64. Native observations remain float32; ridge uses float64 SVD, with slow-band conditioning recorded separately.

Comparators and failures. Equal-weight geometry fits one nonnegative gain to the mean $1 - \cos[\omega_b(d-1)]$. Aligned nonnegative fitting learns one nonnegative weight per band. Shared signed fitting uses mean calibration-query rows instead of each test query's row. The wrong-frequency comparator repeats conditional signed fitting at $1.13\omega_b$. Table 14 retains all grids and doses. Conditional native errors vary substantially; strong positional-numerator prediction does not ensure accurate native-score peaks. For Standard16 at 50% dropout, only 1% of distant maxima are within three lags, although no native distant comparison beats lag 1. Ordinary Base256-16 has 10,805 violating comparisons in the panel and the predictor misses all of them. Zero violations, uncertain locations of weak peaks and token success are distinct outcomes. Full baseline metrics, centered shape errors, wrong-frequency results and seed intervals are retained in the numerical exports.

Table 14: Native independent-content first-layer forecasts on all six grids and three doses (12 seeds per cell). NRMSE compares lag-1 margins at lags 243–480; peak accuracy is the fraction within three integer lags of the native maximum, averaged within model. Neither is copying accuracy.

Grid	Dropout	Equal	Aligned	Shared	Conditional	Peak within 3
Standard4	0%	0.488	0.653	0.065	0.042	0.732
Standard4	25%	0.606	0.774	0.105	0.046	0.897
Standard4	50%	0.625	0.803	0.103	0.041	0.967
Standard8	0%	0.291	0.872	0.095	0.119	0.876
Standard8	25%	0.304	0.943	0.178	0.184	0.319
Standard8	50%	0.360	0.934	0.208	0.213	0.079
Standard16	0%	0.285	0.466	0.235	0.210	0.686
Standard16	25%	0.182	0.278	0.201	0.219	0.339
Standard16	50%	0.328	0.254	0.302	0.313	0.010
Base256-16	0%	0.932	0.907	0.609	0.580	0.385
Base256-16	25%	0.522	0.483	0.228	0.222	0.527
Base256-16	50%	0.484	0.435	0.239	0.211	0.492
Geometric8	0%	0.378	0.291	0.331	0.066	0.884
Geometric8	25%	0.473	0.439	0.393	0.089	0.954
Geometric8	50%	0.675	0.652	0.580	0.225	0.651
Geometric16	0%	0.562	0.598	0.355	0.324	0.383
Geometric16	25%	0.380	0.413	0.200	0.158	0.647
Geometric16	50%	0.389	0.422	0.263	0.166	0.660

N.2 From predicted comparisons to actual outputs

An exploratory join of the frozen forecasts to existing full-copying outcomes tests the rule that an increased mean minimum distant margin predicts improved accuracy. At 25% dropout, predicted margin-change directions agree with native margin changes in 70/72 pairs, but with L240 teacher-forced accuracy changes in 58/72 and greedy token changes in 33/72. Always-increase baselines give 61/72, 49/72 and 24/72, respectively. Spearman correlations of predicted margin change with teacher-forced and greedy change are only 0.164 and 0.097. The 72 pairs share twelve seed identifiers and six grids; they are not 72 independent replicates. These associations neither calibrate copying accuracy nor establish a free-running predictor. L480 comparisons are secondary because they include lags outside the forecasted domain.

Original frozen-selection intervention. At L240 output offset 120, frozen conditional predictions select the twelve strongest distant content keys per head, among lags 243–360. Removing their attention and redistributing it to near content changes only the first-layer current output-query row. Marker/self weights, native values and every other row stay fixed. Eight IID sources crossed with eight nonwrapping origins give 64 outputs per model. Across ordinary models, accuracy changes by $+3.56$ points [2.15, 4.90], versus $+0.50$ for random twelve-key removal and $+0.04$ for predicted bottom twelve. Strengthening the selected keys gives -2.43 points. Learned, random and bottom selections remove mean attention masses 0.0608, 0.0158 and 0.0021: the original controls match key count, not intervention dose. The following replication directly addresses that gap.

Fresh inputs and three output positions. Before new outcomes, we freeze selectors, input seeds and analysis choices for all 216 endpoints. At each of output offsets 60, 120 and 180, 32 fresh IID sources are crossed with four nonwrapping origins, giving 384 outputs per model. Absolute query coordinates are chosen from the already frozen forecast panel and keep the entire L240 sequence inside the 1,024-position domain. Sources are shared across models and origins within an offset, and independently drawn between offsets. Only one current query is edited per execution; the prefix is correct. Geometry-only and wrong-frequency selectors use their frozen short fits and choose twelve keys, exactly as the learned selector does.

For native probabilities p_j, let N be near content (lags at most 242), F distant content, and $T \subset F$ the selected twelve keys. Writing $a_T = \sum_{j \in T} p_j$ and $a_N = \sum_{j \in N} p_j$, the removal is

$$\delta p_j = -p_j \mathbf{1}[j \in T] + a_T \frac{p_j \mathbf{1}[j \in N]}{a_N}. \tag{18}$$

All markers and the current token are outside $N \cup F$. The edit preserves their weights exactly and changes the attention output by $\sum_j \delta p_j v_j$, added to the native attention result. Boosting moves $\min(a_T, a_N/2)$ in the opposite direction; it is not a symmetric dose.

Equal-mass controls. The diffuse control removes exactly a_T proportionally from all F instead of only T. It adds the same vector contribution from near content, isolating which distant values supplied the removed mass. A second pair compares learned and random twelve-key sets at common dose $\min(a_T, a_R)$, removing each set's mass proportionally. Both pairs match the dose exactly for each input and head. Key ranking remains short-only, while intervention amplitude uses native attention; the control is a causal diagnostic, not a deployable short-only repair rule.

Table 15: Fresh-input behavioral effects, averaged across six grids within each seed. All 216 models; 384 correct-history L240 outputs per model. Paired 95% seed-bootstrap intervals; no multiplicity adjustment. Positive CE reduction means lower loss.

Dropout	Intervention	Accuracy gain (pp)	CE reduction (nats)
0%	Learned 12	+3.78 [+2.94, +4.54]	+0.48 [+0.38, +0.58]
0%	Geometry 12	+1.68 [+1.27, +2.07]	+0.23 [+0.17, +0.29]
0%	Wrong frequency 12	+0.59 [+0.43, +0.77]	+0.07 [+0.05, +0.10]
0%	Random 12	+0.84 [+0.59, +1.06]	+0.12 [+0.10, +0.13]
0%	Bottom 12	+0.07 [+0.00, +0.13]	+0.01 [-0.00, +0.01]
0%	Diffuse, same mass	+3.77 [+2.90, +4.54]	+0.47 [+0.37, +0.57]
0%	Matched learned 12	+0.94 [+0.69, +1.18]	+0.13 [+0.11, +0.15]
0%	Matched random 12	+0.83 [+0.58, +1.05]	+0.11 [+0.09, +0.13]
0%	Boost learned 12	-2.47 [-3.18, -1.79]	-0.31 [-0.40, -0.23]
25%	Learned 12	+0.55 [+0.31, +0.77]	+0.03 [+0.01, +0.04]
25%	Random 12	+0.18 [+0.08, +0.28]	+0.01 [+0.00, +0.01]
25%	Diffuse, same mass	+0.64 [+0.41, +0.87]	+0.03 [+0.02, +0.05]
25%	Boost learned 12	-0.54 [-0.72, -0.37]	-0.05 [-0.07, -0.03]
50%	Learned 12	+0.18 [-0.04, +0.43]	-0.01 [-0.06, +0.03]
50%	Random 12	+0.25 [+0.15, +0.36]	+0.02 [+0.00, +0.03]
50%	Diffuse, same mass	+0.37 [+0.26, +0.47]	+0.04 [+0.03, +0.05]
50%	Boost learned 12	-0.29 [-0.53, +0.00]	-0.03 [-0.09, +0.04]

N.3 What the mass controls establish

The fresh-input ordinary-model learned selector improves accuracy by 3.78 points [2.94, 4.54] and reduces CE by 0.480 nats [0.378, 0.576]. Its paired accuracy advantage is 2.94 points [2.18, 3.64] over random selection, 2.10 [1.62, 2.53] over geometry-only selection, and 3.18 [2.45, 3.87] over the wrong-frequency fit. Thus learned comparisons select more consequential distant attention at a fixed twelve-key budget.

The diffuse equal-mass control gives 3.77 points [2.90, 4.54] of rescue. Learned minus diffuse removal is 0.01 points $[-0.16, 0.20]$ and 0.009 nats $[-0.019, 0.037]$. This is no resolved advantage, not a formal equivalence test. At the smaller common dose, learned versus random twelve-key removal has a modest CE advantage, 0.018 nats [0.008, 0.030], but unresolved accuracy advantage, 0.11 points $[-0.03, 0.25]$. Exact key identity therefore does not explain the large original advantage independently of transferred mass. The results support forecast-guided identification of a distant-attention pathway, not uniquely harmful aliases or a complete error-location predictor.

With 25% dropout the learned-selector gain shrinks to 0.55 points [0.31, 0.77]; at 50% it is 0.18 $[-0.04, 0.43]$, with unresolved CE change. Equal-mass diffuse removal slightly outperforms learned removal at 25%. Table 16 retains all adverse and weak cells; the numerical release additionally reports each offset and every paired seed. These are effects of query-row attention edits, not a uniquely mediated fraction of the dropout training effect. The endpoint competition transfers in Appendix L remain complementary tests using native test-length donor information.

Uncertainty and implementation checks. Every statistic averages sources, origins and output offsets within model, then grids within a shared seed for pooled results. Intervals resample the twelve seed means, conditional on the fixed input panels; no tokens or heads are counted as independent training replicates, and no multiplicity adjustment is made. Every native/identity output and past second-layer QKV row is bitwise unchanged. Parameters are compared with checkpoint state after the probe. Probability conservation, nonnegativity, marker/self preservation, exact per-head dose equality and a separate full probability-weighted value reconstruction verify the intervention. Checkpoint, selector and result hashes are checked for all 216 models. The original v1 independent reconstruction also reproduces all 13,824 edited token decisions, with maximum logit difference 0.00574 from native float32 versus double accumulation. These checks concern the measurement implementation, not replication of training.

Table 16: All grids and doses in the fresh-input replication. Learned-selector effects average three output offsets and four origins within model. Equal-mass contrasts subtract diffuse removal from learned-key removal; positive CE benefit favors the learned keys. Paired seed intervals are retained in the numerical release.

Grid	Dropout	Accuracy gain (pp)	CE reduction	Equal-mass CE benefit
Standard4	0%	-0.09	-0.007	-0.000
Standard4	25%	+0.00	-0.000	-0.000
Standard4	50%	-0.02	-0.000	+0.000
Standard8	0%	-0.11	-0.000	-0.001
Standard8	25%	+0.00	+0.000	+0.000
Standard8	50%	+0.00	+0.000	-0.000
Standard16	0%	+4.28	+0.454	+0.022
Standard16	25%	+0.00	+0.000	+0.000
Standard16	50%	+0.00	-0.000	+0.000
Base256-16	0%	+4.56	+0.683	+0.056
Base256-16	25%	-0.26	+0.001	+0.002
Base256-16	50%	+0.20	-0.003	-0.024
Geometric8	0%	+2.06	+0.222	-0.039
Geometric8	25%	+0.67	+0.025	-0.008
Geometric8	50%	-0.87	-0.149	-0.187
Geometric16	0%	+11.96	+1.527	+0.015
Geometric16	25%	+2.89	+0.145	-0.023
Geometric16	50%	+1.80	+0.109	-0.066

N.4 Supporting local-context readouts retained from Figure 3

The original local/distant comparison figure also displayed context response and vocabulary-dependent routing (Figure 21). At fixed 16-band frequencies, source attention on lags 0–7 is 45.9%, 75.8% and 94.0% for $V = 2, 4, 8$. This complements agreement across grids at fixed vocabulary, without specifying a unique local filter. Appendix L retains the alignment model and the history-supervision control rejecting finite-history exposure alone as an explanation of localization.

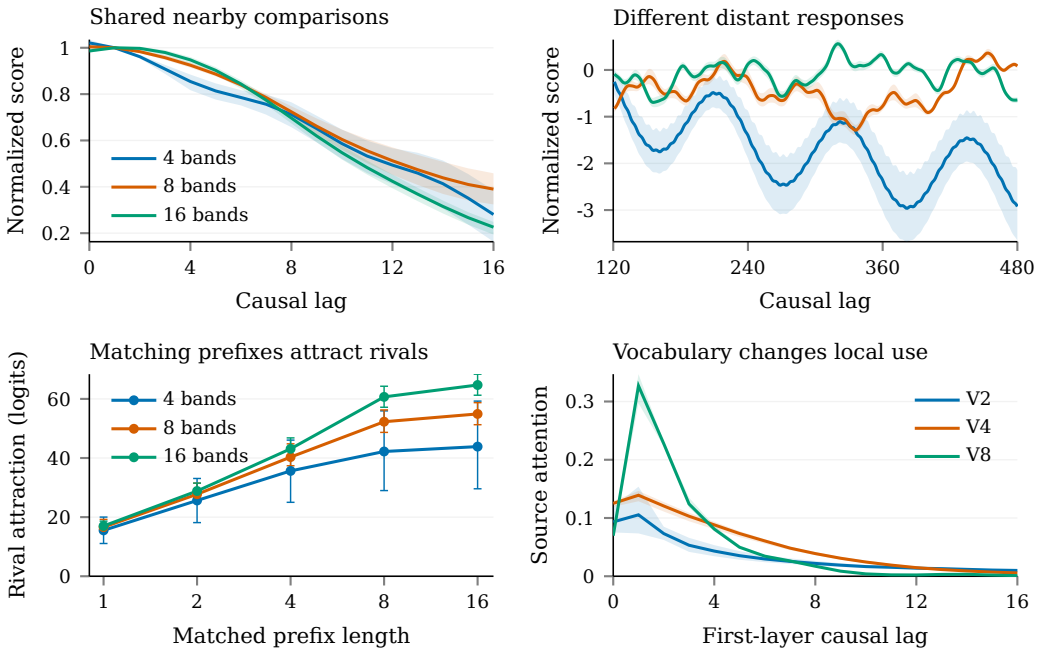


Figure 21: **Local comparisons, context responses and vocabulary-dependent routing.** Top: stationary positional scores, normalized per seed by $K(1)$, on nearby and distant lags. Bottom left: native context response at L120. Bottom right: first-layer source attention with 16 bands and three vocabularies. The band sweep fixes width, frequency endpoints and $V = 4$. Twelve seeds per condition; pointwise 95% seed-bootstrap intervals. The top panels also appear in the revised main Figure 3; they are reused evidence.

O PRETRAINED SPAN-QA: POSITIONAL COMPARISONS AND CONTEXTUAL REPRESENTATIONS

These probes test whether the learned-comparison analysis supplies useful supporting evidence in pretrained models. They separate prediction of an internal comparison, a localized positional intervention, and restoration of context-dependent representations; these are distinct endpoints.

Models, task and evaluation. We use two public 119M-parameter XLNet-Base extractive-QA checkpoints (Yang et al., 2019), with native relative sinusoidal positions: SQuAD2 (revision `36d0bd03`) and SQuAD/Quoref (`1ab9e659`). Both retain 12 layers, 12 heads and width 768; no weights are trained. Controlled clauses associate two subjects with two cities, and a question asks for one subject's city. Swapping the source subjects preserves city tokens, their positions and the question while reversing the correct answer. Eight pairings, two layouts (early or spread across the source), and both assignments give 32 cases per checkpoint and length, without filtering. Neutral repeated filler extends sources from 256 to 4096 or 6144 tokens; question and special tokens add a small overhead. Evaluation uses native context-first QA formatting, no truncation or recurrent memory, and the native five-start by five-end span beam. All questions are answerable; no no-answer threshold is fitted. Answer margin is the correct minus rival city's joint start/end log probability. Confidence intervals resample the eight pair templates, retaining their layouts and assignments together.

Short-context localization. An initial 256-token screen ranks all 144 heads by ablation on four pairings and checks them on four further pairings. L3H8 (one-based indices) is most consequential in both checkpoints. A separate short row screen shows that source-name query rows matter more than the question-name row: validation answer-margin reductions are 1.754/3.176 versus $0.154/-0.091$ nats (SQuAD2/Quoref). The long panel uses eight subsequent pairings; names overlap, so these are not unseen-vocabulary tests. No long intervention gain selects the head or rows. Ablation establishes the importance of these outputs, not the unique importance of a particular key in their attention distributions.

Frozen positional-response prediction. Let $D(q) = S(q, k_{\text{question}}) - S(q, q)$ compare the question-name key with the self key at the correct source-name query. Write its positional score as $b(q, d) = c(q)^\top \phi(d)$, where ϕ concatenates the native sinusoidal features and $c(q)$ includes the query, relative bias, learned projection and attention scaling. For $h(d) = \text{clip}(d, -512, 512)$, the short-query forecast of the clamp response is

$$\widehat{\Delta D} = c(q_{256})^\top [\phi(h(d)) - \phi(d)], \qquad d = q - k_{\text{question}}. \tag{19}$$

The self lag remains zero and content/segment terms cancel in this intervention difference. This reads the learned projection and short queries directly, a stronger white-box information budget than the passive fits in Appendix N. Native long queries can differ. Predictions are frozen before the row-specific interventions; native and whole-head outcomes on this panel were already known, making this a targeted follow-up rather than an independent prospective replication.

Prediction and actual behavior. At 4096 tokens, response correlations are 0.840/0.955 and RMSEs are 0.312/0.293 (Figure 22). The intervention recomputes only the two source-name query rows of L3H8, clamping the positional term while retaining native long Q/K/V, every eligible key and normal downstream computation. In SQuAD2, answer margin rises by 0.375 nats [0.241, 0.518], recovering 79% of the whole-head margin gain; span exact match rises from 78.125% to 90.625%. Quoref gains 0.097 nats [0.046, 0.158], but exact match falls from 59.375% to 56.250%. Question-row edits give $-0.011/-0.014$ nats; the same source-row edit in L3H9 gives $+0.010/-0.001$ nats. The edit changes the entire positional score row, so D is a predictive diagnostic, not an isolated causal mediator. Predicted score changes correlate descriptively with answer-margin gains ($\rho = 0.578/0.680$), without calibrating final QA behavior.

Absolute-state prediction is weaker. Retaining short content/segment scores and mapping long keys by token, segment, clause role and filler phase gives full-comparison RMSE 2.873/0.970, versus 2.541/1.832 when retaining the short measurement unchanged. In SQuAD2, predicted question attention mass averages 0.258 versus the native 0.737. Thus high score-rank correlation does not establish calibrated probabilities or prediction of native mistakes.

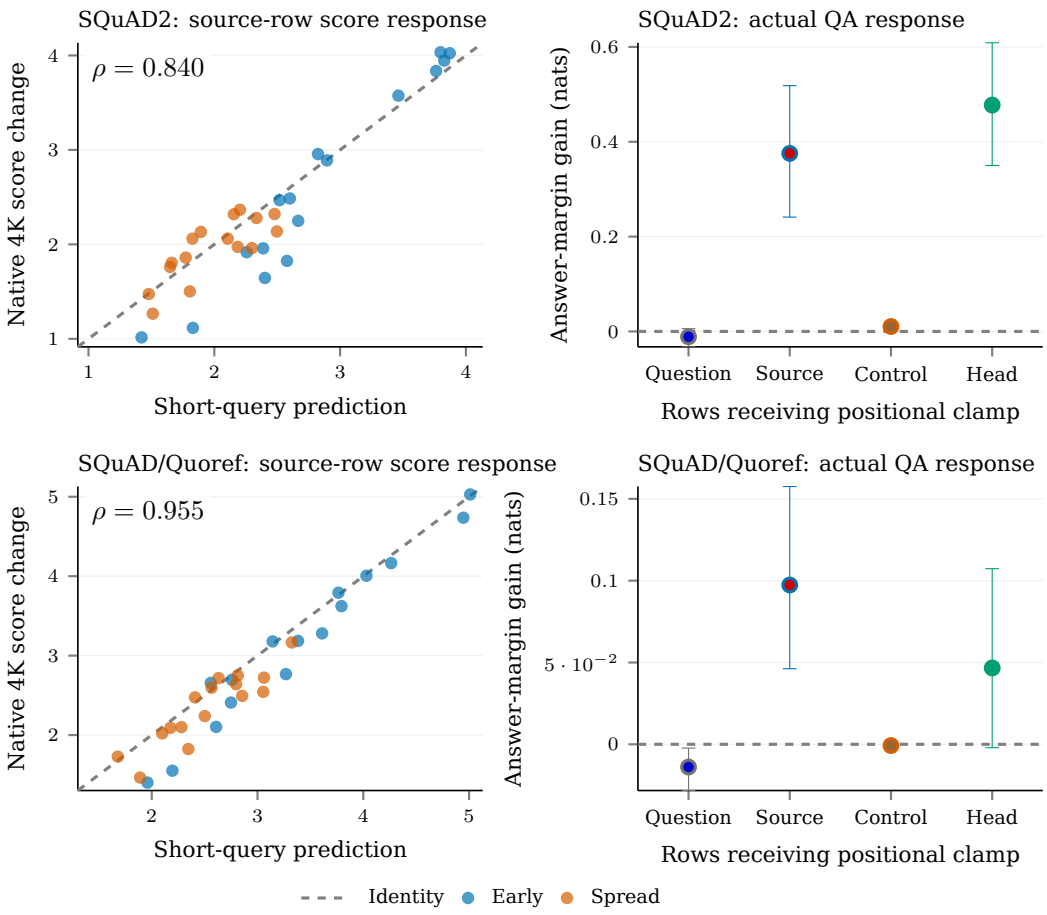


Figure 22: **Short-query predictions and localized pretrained QA effects.** Left: predicted versus native 4K change in the source-row comparison under relative-lag clamping; points cover both assignments and early/spread layouts. Right: final answer-margin changes when clamping question-name rows, source-name rows, the same source rows in control head L3H9, or all L3H8 rows. L3H8 and source rows were selected on short examples. Intervals are 95% pair-template bootstraps. Score and answer changes are separate endpoints.

Attention weights and contextual values. A separate short-context screen selects four heads per checkpoint by the effect of opposite-assignment donor outputs at the two city tokens: L10H9/L10H4/L9H7/L8H1 for SQuAD2 and L10H4/L10H9/L11H11/L11H3 for Quoref. At 6144 tokens, let A_L, V_L be native attention and values, and A_P, V_S the attention forecast and mapped values from short states, preserving token, segment and task role while including all long keys. We replace only the selected city-token head outputs. Mixed products are computed from one unmodified long forward before grouped replacement; downstream computation then propagates normally. This exploratory decomposition was designed after observing earlier 2K restoration results.

Both weights and values contribute (Table 17). Opposite-assignment short-output donors give zero exact match despite fixed city tokens, positions and question, making the rescue association-dependent. These conditions use privileged donor states, and mixed conditions also use native long states; they are diagnostics, not deployment methods. Values already contain upstream positional and content processing, so their rescue does not isolate the original cause or establish orthogonal mechanisms. The earlier short-only forecast evaluated through the short QA readout predicted restoration gains of only $+0.111/-0.057$ nats, versus actual short-output gains of $+7.705/+5.675$; successful oracle restoration does not validate that failed end-task forecast.

Table 17: Selected city-head interventions at 6144 source tokens. EM is span exact match (%); Δm is mean answer-margin gain (nats). All 32 cases per checkpoint, grouped over eight pair templates. Outputs change only at the two city tokens in four short-selected heads.

		SQuAD2		Quoref	
Edit	Output	EM	Δm	EM	Δm
Native	$A_L V_L$	81.2	+0.000	53.1	+0.000
Attention only	$A_P V_L$	93.8	+2.141	84.4	+1.609
Values only	$A_L V_S$	100.0	+1.886	87.5	+3.655
Both	$A_P V_S$	100.0	+7.277	100.0	+5.754
Short output	Short donor	100.0	+7.705	100.0	+5.675
Swapped output	Opposite assignment	0.0	-11.853	0.0	-6.313

Numerical checks and scope. Native CPU float32 sinusoidal features are preserved; coefficient diagnostics use float64. Identity and short-control spans are unchanged. Independent captures verify that source-only edited outputs match the corresponding whole-head clamp within 6×10^{-6}, while every other row/head output at the edited layer is bitwise unchanged. Native Q is unchanged there. The inputs exceed the published XLNet-Large reference recipe's 512-token chunk plus 384-token memory[1], but exact community fine-tuning length caps are undocumented. Two related fine-tunings and synthetic filler do not establish a natural-document or architecture-wide result. The evidence supports a limited pretrained extension of learned-comparison analysis, not geometric prediction of final QA errors, identification of the toy prefix-matching circuit, or mediation of dropout gains by compensation. Checkpoint hashes, frozen predictions, all outcomes and numerical audits accompany the released analysis.

[1] https://github.com/zihangdai/xlnet