ENDPOINT RELATIONS EXTEND QUANTITATIVE READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reading familiar quantities is different from extending their scale. We study whether relations anchored to observed values can improve predictions beyond the range used to train a language-model readout. We learn a relation between distance from a source endpoint and nearby values, then use it to correct a global reader without new-range labels. Across four ordered tasks and 48 checkpoints, normalized absolute error decreases from 0.714 to 0.632, an 11.45% reduction, after selecting the global reader by its final prediction loss. The intact relation also outperforms controls that preserve the endpoint while disrupting the other pairings. Comparisons using the same features locate the useful ingredient: fitting the radius on all source values reduces error further to 0.529, and simple direct readers capture much of the gain. Controlled metric experiments then separate accessible quantity information from a distance that organizes it and a relation that continues across ranges. Together, these results show that quantitative extrapolation depends on the relation used to read a representation, and that endpoint relations can expose variation a global reader leaves unused.

1 INTRODUCTION

Reading a scale is not the same as extending it. A language-model representation may distinguish familiar quantities while a reader trained on those quantities compresses their spacing or misorders larger values. Training a new reader on the missing range can reveal that the information remains. The more useful question is what the original observations already tell us about that range, before its labels arrive.

An observed endpoint supplies a natural reference. A scalar projection asks how far a representation moves along one fitted direction. Distance from the endpoint asks how far it moves from a known observation. These measurements can vary differently, even on the same frozen representation. If endpoint distance tracks numerical change in the source, it supplies a second coordinate from which to extend the reading.

We make this idea explicit with an endpoint correction. A small regression relates endpoint distance to source values near the boundary. Its disagreement with a global prediction is detrended on the source and added back as a fixed correction. All transformations are fitted on source data; the direction of extrapolation is supplied; each query is predicted separately. Figure 1 shows the complete information flow.

The central finding is that source-derived relations expose quantitative variation that a fitted global reader leaves unused. Across 48 checkpoints and four short ordered tasks, the correction reduces normalized absolute error by 11.45%, improves 358 of 384 checkpoint–task–direction conditions, and improves both ordering and pairwise changes. This result uses a global pipeline selected for the final prediction loss. An endpoint-preserving control further isolates a contribution from the correspondence among the other source observations.

The same comparisons sharpen the meaning of "local." The reference is local—an observed boundary point—but the relation need not be fitted only near that point. Fitting the same radius on all source observations improves accuracy further. A simple source-selected reader also uses this information effectively. Thus the evidence identifies a useful measurement coordinate, while separating that coordinate from the particular correction used to expose it.

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.

A matched geometry study tests which aspects of a quantitative relation continue across ranges. Entity attributes are recoverable from the full representation even when endpoint distance is a poor summary. Learning an attribute-related distance improves its alignment inside the source but does not preserve that alignment in the held range. Information availability, metric alignment, and continuation are therefore distinct properties of a quantitative reading.

Our contribution is a controlled account of how these properties interact: an executable source-only correction; paired tests of reader selection, fitting support, and endpoint correspondence; and interventions that change the metric while holding the underlying observations fixed. The result is a constructive view of quantitative extrapolation as learning a relation that remains useful beyond the range on which it was fitted.

2 Related Work

What probes measure. Probes expose information recoverable from frozen representations (Alain & Bengio, 2017; Belinkov, 2022). Control tasks and description-length analyses clarify the role of the decoder's capacity (Hewitt & Liang, 2019; Voita & Titov, 2020). We ask a complementary question: whether a reader trained on one numerical range preserves values, ordering, and changes on another. Recoverability within a range and reuse across ranges require different evaluations.

Quantities in representation space. Language models encode numerical magnitudes, spatial and temporal coordinates, and monotonic entity attributes (Wallace et al., 2019; Gurnee & Tegmark, 2024; Heinzerling & Inui, 2024). Representation geometry also depends on the measurement and comparison operations (Park et al., 2024), and some features are intrinsically multidimensional (Engels et al., 2025). These findings motivate additional coordinates beyond a fitted projection. A quantity may be recoverable through one reader while poorly aligned with another. We test which source-fitted reading relation continues beyond its observed range.

Local regression and extrapolation. Local fitting and residual addition are established statistical tools (Cleveland, 1979; Friedman, 2001). Our correction uses them to test the incremental value of an endpoint-relative feature. Blocked evaluation exposes shifts hidden by random splits (Roberts et al., 2017), and model selection must include the transformations used at evaluation (Cawley & Talbot, 2010). We combine these principles with endpoint-preserving controls and same-feature alternatives to distinguish a helpful relation from a particular choice of estimator.

3 Method: Extend a Reading from an Observed Endpoint

3.1 Source supervision and a fixed global reader

Let $S = \{(h_i, y_i)\}_{i=1}^{n_S}$ contain frozen representations and source labels. Query values lie below or above the source range. Every method receives the direction $d \in \{-1, +1\}$, and the endpoint e maximizes dy_i over S. Lower and upper readers are fitted separately. Query representations are available for prediction; their labels, ranks, and batch statistics are not used for fitting. A value is "unseen" here when excluded from probe supervision, irrespective of language-model pretraining.

The global branch fits source PCA and ridge, followed by a source-fitted scalar continuation, to produce $q_{g,S}$. Its final prediction q_S calibrates displacement from the endpoint:

$$v_i = |q_{g,S}(h_i) - y_e|, \qquad c_S = \frac{\sum_{i \in S} v_i |y_i - y_e|}{\max\{\sum_{i \in S} v_i^2, 10^{-12}\}}, \qquad q_S(h) = y_e + d\, c_S |q_{g,S}(h) - y_e|. \quad (1)$$

To select the reader, we temporarily hold out a boundary of the source and refit every stage on the remaining inner source. The main analysis minimizes the final q_S validation loss over 4,800 nominal configurations. We also compare selection by the upstream $q_{g,S}$ loss. The two policies are compared separately; Appendix A specifies the grid and tie handling.

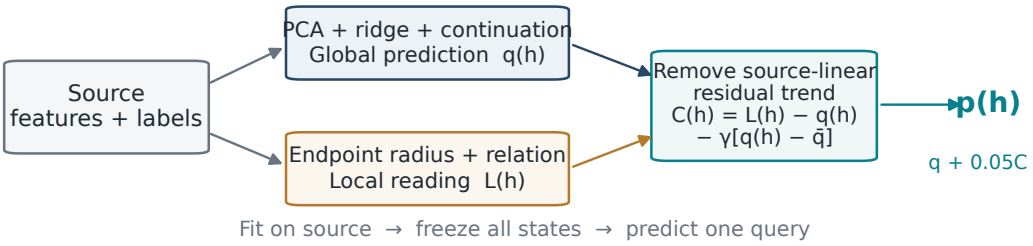

Figure 1: **Use a known endpoint to obtain a second reading.** The global branch compresses source features through PCA; the radial branch uses the full source-standardized layer. Both fit the same source labels. Source predictions determine the residual trend. The resulting states are frozen, and the two branches predict each query independently. The second coordinate measures displacement from an observed reference.

3.2 An endpoint-relative relation

Both branches use stored depth fraction 0.75. Let z_S standardize each hidden coordinate by its source mean and population standard deviation, replacing deviations below 10^{-8} by one. Define

$$\rho_S(h) = \|z_S(h) - z_S(h_e)\|_2, \qquad u_S(h) = \frac{\rho_S(h) - \bar{\rho}_{B_S}}{\max\{\mathrm{sd}_{B_S}(\rho_S), 10^{-8}\}}, \tag{2}$$

where B_S contains the extreme $k = \max\{3, \min(n_S, \mathrm{round}(0.4n_S))\}$ source items in direction d. "Local" refers to this boundary fitting set; it is selected by source target order. A two-parameter ridge fit gives the relation and its outward continuation:

$$(\hat{b}, \hat{a}) = \arg\min_{b,a} \sum_{i \in B_S} [y_i - b - a u_S(h_i)]^2 + 10^{-3} a^2, \qquad L_S(h) = y_e + 3d\,|\hat{b} + \hat{a} u_S(h) - y_e|. \tag{3}$$

The intercept is unpenalized. The endpoint label places the reading on the scale, and the supplied side determines the continuation direction. The multiplier 3, tail fraction 0.4, and gain below are held fixed across the core comparisons.

3.3 Correct the variation left by the global reading

On source items, set $r_i = L_S(h_i) - q_S(h_i)$ and let bars denote source means. Remove the source-linear trend of the residual:

$$\gamma_S = \frac{\sum_{i \in S}(q_S(h_i) - \bar{q}_S)(r_i - \bar{r}_S)}{\max\{\sum_{i \in S}(q_S(h_i) - \bar{q}_S)^2, 10^{-12}\}}, \tag{4}$$

$$C_S(h) = L_S(h) - q_S(h) - \gamma_S[q_S(h) - \bar{q}_S], \qquad p_S(h) = q_S(h) + 0.05 C_S(h). \tag{5}$$

This removes a slope, while retaining the mean disagreement. With the denominator floor inactive, $\mathrm{Cov}_S(q_S, C_S) = 0$ and $\bar{C}_S = \bar{r}_S$. The construction makes the source-linear overlap explicit; its predictive value is measured on held values. The gain 0.05 fixes the coefficient, not an upper bound on the prediction change.

Fit, freeze, predict. Select and refit the global branch on source; fit the radial relation; evaluate both on source to store $\bar{q}_S$ and γ_S. Freeze these states. For each query, evaluate the two branches and return Eq. (5). On a region with fixed reflection sign, this is a constrained linear combination of $q_S(h)$ and $\rho_S(h)$ plus an intercept. It can distinguish queries with equal global scores but different radii. This form also identifies direct radius regression and joint score–radius regression as close alternatives.

When a second reading reduces to an offset. The construction identifies a precise case in which a relation contributes no new query-dependent variation. Suppose $L_S(h) = a q_S(h) + b$ on both source and query observations, with the variance floor inactive. Then $r = (a-1)q_S + b$, so $\gamma_S = a - 1$ and

$$p_S(h) = q_S(h) + 0.05[(a-1)\bar{q}_S + b]. \tag{6}$$

Every query receives the same shift. Two globally affine readers of the same radius satisfy this condition when the first has nonzero slope; reflected or piecewise readers need not. The identity makes complementarity concrete: it depends on what the second reading adds to the first, not simply on whether the representation contains a quantity. The experiments below test that dependence by changing the starting reader while retaining the relation.

3.4 Separate the anchor, relation, and fitting support

An endpoint has a special status: its radius is zero and its target is known. Our *endpoint-preserving control* fixes this pairing and deranges all other tail assignments, refitting both the local line and residual trend. We score every resulting prediction separately and then average losses. A full-tail control also moves the endpoint pairing; an endpoint-indicator control retains only endpoint membership. These are distinct comparisons, rather than interchangeable permutation tests.

To test the fitting support, we replace B_S by all of S while retaining the radius, multiplier, baseline, and correction rule. To test the reader, four direct ridge families cross radius versus full hidden features with boundary versus all-source fitting. Each chooses eight penalties and direct versus endpoint-reflected output on the same source-pseudo split. Selecting the best family gives a 64-candidate simple reader s_S. Its output class and search budget differ from the global pipeline; all methods use the same source labels, depth, and direction. We also apply the unchanged correction and endpoint-preserving control to s_S.

4 Experiments and Results

4.1 Range holdouts and evaluation

The core contains number words, ordinals, percentages, and probability expressions: 70 items, 48 checkpoint IDs, and two held boundaries, giving 384 conditions. Table 1 shows the actual fitting and evaluation sizes. Probability expressions use assigned ordinal codes. The local configuration was selected using source-pseudo evaluations on 17 other task identities; the four core identities were excluded from that selection roster. The controlled comparisons share the same core panel.

Table 1: **The core tests continuation beyond labeled values.** Counts list lower/upper holdouts. The three numerical lists contain 59 items. Probability codes order expressions but are not calibrated probability units.

Task	Items	Source	Held	Boundary fit
Number words	20	14/12	6/8	6/5
Ordinals	20	14/12	6/8	6/5
Percentages	19	13/12	6/7	5/5
Probability expressions	11	8/6	3/5	3/3

The primary metric is held MAE divided by the population standard deviation of source targets (nMAE). All-pairs change error measures differences between predictions in the same units; Spearman measures ordering; unit error is $|a - 1|$ for prediction-on-target slope a. We average conditions equally. Paired bootstrap intervals average directions before resampling checkpoints and task identities (Owen, 2007; Owen & Eckles, 2012). We use 20,000 draws, report descriptive crossed 95% intervals, and provide task-fixed and three-comparison adjusted intervals in Appendix C. The four shared task lists and related checkpoints determine the scope of these intervals.

4.2 A relation improves the complete selected reader

Selecting by final prediction loss gives global nMAE 0.714111 and corrected nMAE 0.632372: an **11.45% reduction**. The paired difference is -0.081740, with crossed 95% interval $[-0.104158, -0.058133]$. Improvement occurs in 358/384 conditions and in the average of every core task and both directions. Excluding the probability codes retains the numerical-task effect (Appendix B).

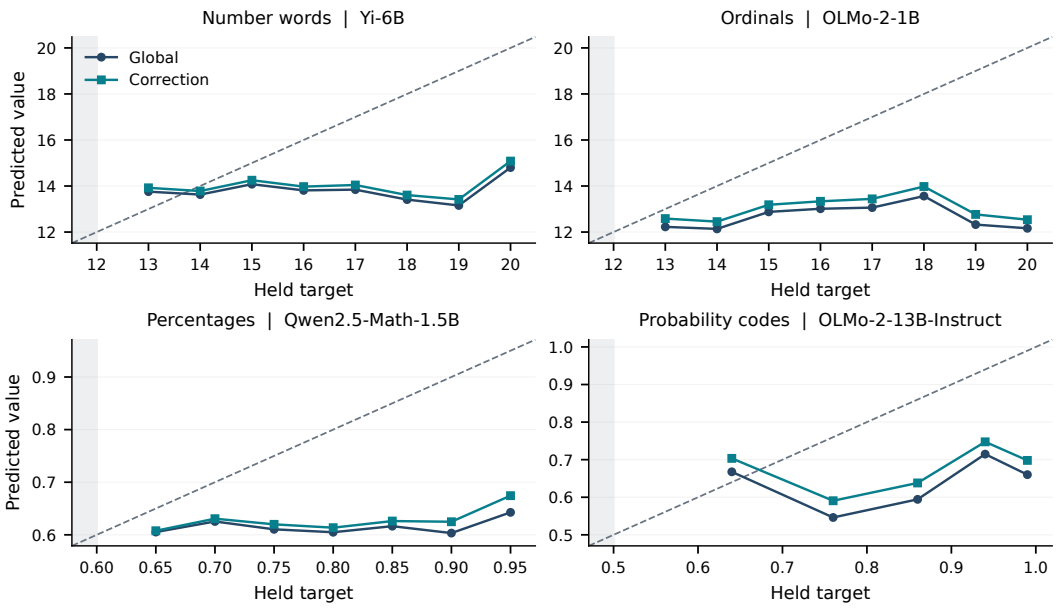

Figure 2: **How the correction changes unseen values.** Actual upper-range query predictions, in native target coordinates. For each task, the displayed checkpoint is closest to the median paired nMAE improvement among all 48 checkpoints; ties use checkpoint name. This descriptive selection is applied after evaluation and does not fit the method. Dashed diagonals denote exact readings. The shaded edge marks the source endpoint. All conditions and the selection data are supplied in the supplement.

The improvement changes the reading, rather than just its location. Change error falls from 0.626726 to 0.592321, Spearman rises from 0.382986 to 0.493477, and unit error decreases from 0.810781 to 0.745332. A source-output offset reaches nMAE 0.684200 while preserving the baseline's ordering. Figure 2 shows actual query predictions selected by a common, outcome-transparent rule across the four tasks.

Upstream-loss selection gives a closely aligned result, $0.718854 \rightarrow 0.634525$, or 11.73%. Thus selecting the complete baseline preserves the correction's benefit. These are two selections on one feature cohort, with all 384 conditions retained under each policy.

4.3 Correspondence matters beyond the endpoint

Keeping the endpoint pairing intact while deranging the others gives nMAE 0.645342, compared with 0.632372 for the intact relation. Their difference is -0.012970 $[-0.024726, -0.004192]$, favorable in 288/384 conditions. Each task has a favorable mean. The endpoint supplies an anchor, but does not account for all of the correspondence-specific benefit in this pipeline.

The control hierarchy helps locate that benefit. Full-tail disruption gives 0.724073; an endpoint-indicator correction gives 0.676309. The smaller intact-versus-endpoint-preserving contrast measures what remains once the privileged anchor is retained. It should be interpreted as an incremental comparison, not as a decomposition assigning the whole 11.45% reduction to nonendpoint geometry. Table 2 places these contrasts beside the closest same-feature alternatives.

4.4 A local reference does not require a narrow fitting set

Fitting the same radius on all source values lowers corrected nMAE to 0.528889. Nothing about the endpoint, global predictions, multiplier, or gain changes. Broadening the support strengthens this reading. The distinction is useful: choosing a reference point and choosing how many observations estimate its relation are separate decisions.

Table 2: **The useful coordinate can be read in more than one way.** Means over the same 384 conditions. The first block uses the final-loss-selected global branch. "All-source radius" changes the fitting support within the correction. The second block contains direct source-selected readers and the unchanged correction applied to the selected reader s_S. Null losses are averaged over individual assignments.

Reader	nMAE ↓	Change ↓	Rank ↑	Unit ↓
Global q_S	0.7141	0.6267	0.3830	0.8108
Endpoint correction p_S	0.6324	0.5923	0.4935	0.7453
Endpoint-preserving control	0.6453	0.6015	0.4778	0.7606
All-source radius correction	0.5289	0.5554	0.5711	0.6680
Joint score–radius correction	0.6157	0.6205	0.3999	0.7947
Standalone boundary-radius ridge	0.5349	0.6048	0.7865	0.7724
Selected simple reader s_S	0.5311	0.6169	0.6730	0.7819
s_S + unchanged correction	0.5123	0.6102	0.7034	0.7635
s_S + endpoint-preserving control	0.5115	0.6112	0.6965	0.7680

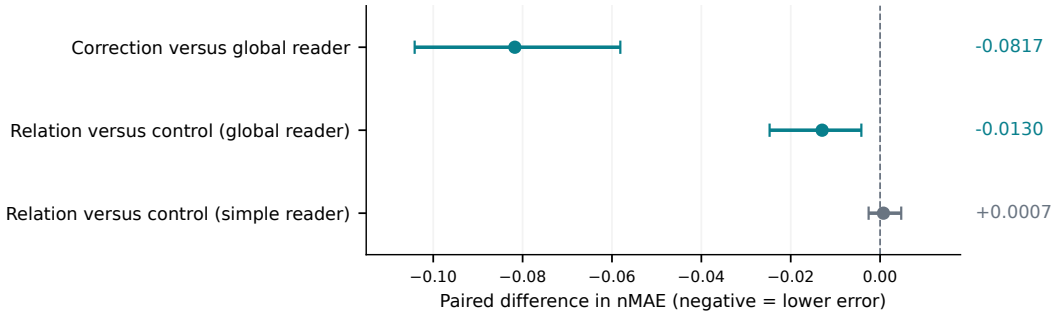

Figure 3: **Complementarity depends on the reader.** Paired differences over the same 384 conditions, with crossed checkpoint–task 95% bootstrap intervals. The relation improves the selected global reader and beats its endpoint-preserving control. For the stronger simple reader, the corresponding intact-versus-control interval includes zero. Each row answers a different question; the contrasts are not additive.

Direct readers reinforce the result. Standalone boundary-radius ridge reaches 0.534938. Selecting among all four simple families gives $s_S = 0.531066$; the chosen family is boundary radius in 259 conditions, all-source features in 81, all-source radius in 35, and boundary features in nine. The endpoint correction has lower mean change and unit errors than s_S, but s_S has lower primary error and better rank. The evidence therefore supports the value of radial information more broadly than the original residual construction.

The value added also depends on the starting reader. Applying the unchanged correction to s_S yields $0.531066 \rightarrow 0.512280$. Its endpoint-preserving control reaches 0.511533, however, with an intact-minus-control difference of $+0.000747$ $[-0.002565, +0.004731]$. The correspondence-specific advantage established for q_S is not resolved for this stronger reader. This identifies complementarity as a property of a *relation–reader pair*: once a reader already exploits the coordinate, its remaining benefit can change. These simple-reader comparisons are exploratory; each condition's selection uses source labels only.

Source validation can also choose how broadly to fit the relation. A separate extension jointly selects tail fraction, displacement multiplier, and gain, reducing core nMAE to 0.541939. It selects the full-source tail in 215/384 conditions, consistent with the fixed-support comparison. This adaptive rule is reported separately from the frozen correction; its complete parameter, layer, and transfer results appear in Appendix D.

Full-feature ridge has access to every hidden coordinate but combines them through a projection. Explicit radius regression instead reads distance from the endpoint. Access to the same raw features therefore does not make these reading rules equivalent. Figure 4 isolates the distinction within the

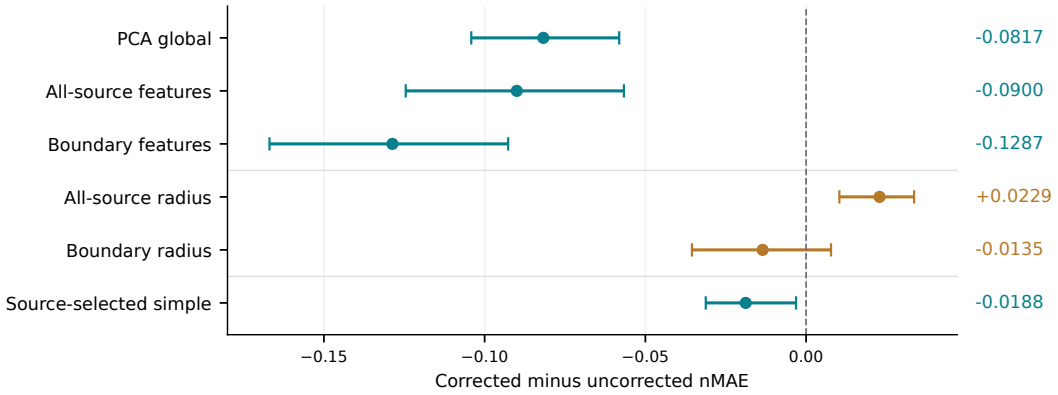

Figure 4: **A radial relation contributes differently to different starting readers.** The boundary relation, displacement multiplier, and gain are fixed; only the starting reader and its source residual trend change. Points are means over the same 384 conditions and bars are crossed 95% intervals. Direct feature readers gain more than direct radial readers. All comparisons reuse one core panel.

Table 3: **The correction aligns less with errors left by a stronger reader.** Exact accounting of the fixed gain-0.05 predictions. Signed movement includes both helpful and harmful directions; the overshoot term is subtracted. No query has an initially zero residual. This retrospective analysis uses held labels to explain saved outcomes, while all fitted predictions remain unchanged.

Starting reader	Signed movement	Overshoot cost	Net nMAE reduction
PCA global q_S	0.087659	0.005919	0.081740
Selected simple s_S	0.025131	0.006345	0.018786

direct families: their source labels, direction, and selection split match, while the feature being fitted changes. This provides a concrete way to ask which information a baseline already uses.

4.5 How the correction changes prediction error

A fixed-prediction analysis connects reader dependence to the direction and size of actual corrections. Let $e = y - q$ and $\delta = p - q = 0.05C$. For a nonzero error, define the signed movement toward the target as $t = \text{sign}(e)\delta$. Then

$$|e| - |e - \delta| = t - 2(t - |e|)_+. \tag{7}$$

Movement toward the target reduces error until the prediction passes it; the second term accounts for that excess. Movement away has negative t. An initially exact prediction instead contributes $-|\delta|$. Averaging these terms within each condition in source units gives an exact accounting of the primary metric (Table 3).

For the global reader, the correction moves toward the target on a mean 83.3% of queries per condition; for the selected simple reader, this is 61.8%. The smaller improvement reflects lower signed alignment with the remaining errors, with a similar overshoot cost. Across the direct families, full-feature readers benefit substantially, whereas correcting the all-source radial reader increases nMAE from 0.731237 to 0.754094. These outcomes agree with the distinction in Eq. (6): access to a coordinate and useful residual variation from that coordinate are separate questions. Appendix A.7 gives the complete gain sensitivity of these frozen predictions; gain 0.05 remains the primary setting.

4.6 New checkpoint extractions and new notations

We ran text-to-feature-to-prediction experiments at fixed revisions of GPT-Neo-125M and Pythia-160M-deduped. Across the same four tasks, final-loss selection gives $0.808885 \rightarrow 0.747152$ nMAE, a 7.63% reduction. These 140 newly extracted text representations provide new checkpoint observations on the established task definitions. A separate pinned Granite-4.0-H-350M-base execution

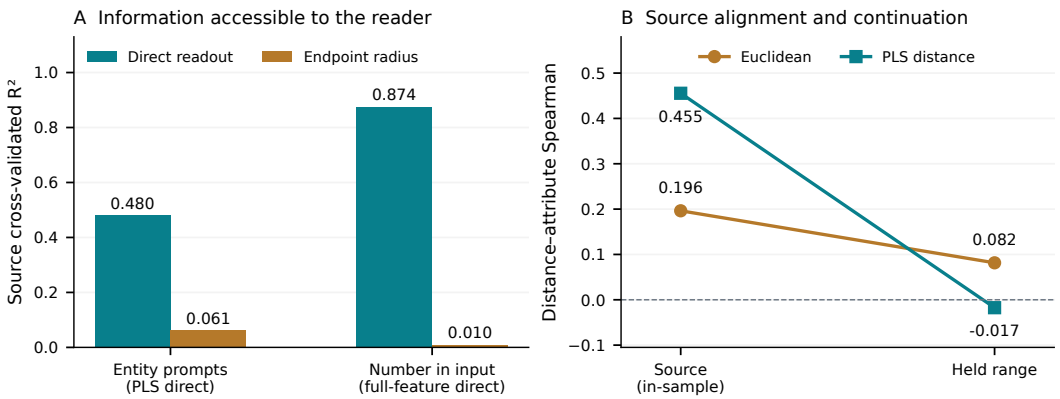

Figure 5: **Decoding a value, organizing a distance, and extending a relation are different tests.** **A:** Random five-fold source R^2. The direct readers are PLS for entity prompts and full-feature ridge for the number-input positive control; every learned transform is refitted inside the fold. **B:** Endpoint-excluded distance–attribute Spearman on original entity inputs. PLS uses source labels, so its source alignment is an in-sample statistic. A larger source association does not persist in the held range. Values average the same two-model, six-attribute, two-direction panel; panels A and B measure different quantities.

adds eight upstream-selection conditions, with nMAE $0.720770 \rightarrow 0.652587$; its method and cohort remain separate.

Roman numerals and duration strings test a different continuation problem: each notation has 128 items, with 102 source and 26 held per side, on both checkpoints. The fixed correction gives $0.530745 \rightarrow 0.537236$ under final-loss selection. Source-selected boundary-feature and boundary-radius readers instead reach 0.222396 and 0.254488. This result directs attention to the measurement used for a notation, rather than treating a single correction as transferable by default. The complete two-policy results appear in Appendix E.

4.7 Information, distance, and continuation separate

We next intervene on the distance itself. Qwen3.5-4B and Qwen2.5-0.5B are newly extracted on the core and on six entity attributes with 512 records each: area, birth year, death year, elevation, inception year, and population. Matched prompts and pooling choices hold entity identities and targets fixed. The main study comprises 224 conditions. Unless stated otherwise, the following comparisons use the original prompts and carrier-span mean pooling.

Information can be accessible through a reader and poorly summarized by a radius. In random five-fold source evaluation, a four-dimensional PLS readout reaches mean $R^2 = 0.4798$ on entity attributes, while endpoint-radius regression reaches 0.0614 on the same representations. A positive control puts the recorded target number directly in the input. Full-feature ridge then reaches $R^2 = 0.8736$, yet radius regression remains at 0.0100 (Fig. 5A). These fold-local source diagnostics show that a weak radius need not indicate missing quantity information.

A more aligned source metric need not extend more faithfully. Holding hidden vectors, global predictions, splits, direction, multiplier, and gain fixed, we replace Euclidean distance with cosine, PCA, random projection, PLS, or source-graph distance. PLS increases source distance–attribute Spearman from 0.1963 to 0.4555, while held association changes from 0.0815 to -0.0175 (Fig. 5B). Source-supervised alignment alone does not supply the missing continuation.

The corresponding entity nMAEs are 1.3564 for the global reader, 1.3522 for Euclidean correction, and 1.3513 for PLS correction. The Euclidean-minus-global difference is -0.0042, with a descriptive family-bootstrap interval of $[-0.0183, +0.0143]$. On the core tasks with these same models and extraction settings, Euclidean correction gives a difference of -0.0509, with interval $[-0.0680, -0.0301]$; PLS is worse than Euclidean by 0.0238. The panel difference therefore persists under matched extraction.

Selecting metric, tail, and gain through nested source boundary holdouts yields entity nMAE 1.3369, a difference of -0.0152 relative to fixed Euclidean correction, with interval $[-0.0308, -0.0033]$. A gate based only on source tail correlation does not outperform gain selection. Source prediction across a temporary boundary is a more direct test of the intended use than the appearance of an aligned source plot. Appendix F reports every metric, prompt/pooling condition, and source-size and spacing intervention.

5 DISCUSSION

The experiments connect local and global readout through a single question: *which source relation remains useful when the numerical range changes?* An endpoint-relative coordinate can complement a fitted global score, and even a small fixed regression can expose that contribution. The stronger all-source and direct readers show that the coordinate and its estimator should be evaluated separately. A local reference can support a broad fit.

This perspective also clarifies the geometry results. If a representation contains $h = (y, u)$, a direct reader can recover y while Euclidean distance also accumulates variation in u. Improving the metric may make y more visible on the source; continuation still requires its relation to distance to persist beyond that source. Our metric interventions test these stages separately, without attributing every failure to curvature or every successful probe to a transferable scale.

This distinction changes how an apparent extrapolation limit should be diagnosed. If a direct reader improves predictions on the same frozen states, the limit belongs partly to the reading rule. If a metric raises source alignment but leaves held-range error unchanged, the missing ingredient is continuation of the learned relation. If an unchanged correction loses its correspondence-specific advantage against a stronger reader, the extra coordinate contributes less beyond that reader. These matched interventions turn a broad question about whether a model represents a quantity into concrete questions about which transformation recovers it, which observations estimate that transformation, and where it remains useful.

For practical readout design, the implication is to validate the complete prediction rule across source boundaries. Preserve the anchor when testing the remaining correspondence, compare readers that use the same features, and evaluate learned distances by predictions across a range gap. A source-aligned metric is a candidate measurement whose continuation must be tested.

Scope. The core consists of four short lists reused across related checkpoints. The later same-feature comparisons and source-readout diagnostics are retrospective; new checkpoint extractions do not create independent task identities. Entity targets are inherited benchmark records, and number-input controls intentionally reveal those targets. Our results establish reader-specific gains and explain observable transfer boundaries; broader task coverage and a reliable source-only criterion for successful continuation remain open. Fitting details and numerical execution are specified in Appendix G.

6 CONCLUSION

Relations anchored to observed values can improve quantitative predictions beyond the supervised range. Endpoint-preserving controls identify useful correspondence beyond the anchor for the tested global reader, while same-feature alternatives show how that information can be used more effectively. Metric interventions reveal the remaining challenge: making an accessible quantity into a relation that continues. **Extending a quantitative readout means extending a measurement relation.**

AI USE STATEMENT

Generative AI tools assisted with research ideas and hypotheses, experimental design, implementation, synthetic diagnostic construction, mathematical reasoning and exposition, result interpretation, literature checking, translation, figure preparation, and manuscript drafting and editing. Computational checks include separate reconstruction of prediction metrics, selected baseline fits, and radial correction formulas, as described in Appendix G. The authors are responsible for the final content, claims, references, artifacts, and submission.

REPRODUCIBILITY STATEMENT

The supplementary materials contain the implementation, source labels and separate evaluation targets, selected configurations, component predictions, individual pairing-control predictions, and scripts for reconstructing the reported metrics and figures. Appendix A specifies the fitting and selection procedures; Appendix G records extraction provenance, numerical checks, and execution precision. Data supporting the main tables and figures are included in machine-readable form.

ETHICS STATEMENT

The experiments analyze existing model representations and benchmark records. Assigned probability-expression codes are ordinal measurements rather than calibrated judgments about people. Entity labels retain the source benchmark's transformations and are not a verified current factual database. Models and data remain subject to their original licenses.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop*, 2017. URL https://arxiv.org/abs/1610.01644.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979. doi: 10.1080/01621459.1979.10481038.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2405.14860.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 175–195. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.18. URL https://aclanthology.org/2024.acl-short.18/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122. URL https://arxiv.org/abs/0712.1111.

Art B. Owen and Dean Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012. doi: 10.1214/12-AOAS547. URL https://arxiv.org/abs/1106.2125.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of EMNLP*, pp. 183–196, 2020.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5307–5315, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534/.

A Fitting and Selection

A.1 Task wording and assigned values

Number words are one through twenty, with targets 1–20 and the template `the number is {item}`. Ordinals are first through twentieth, with targets 1–20 and `they finished in {item} place`. Percentages are five percent through ninety-five percent in steps of five, with targets 0.05–0.95 and `approximately {item} of the population`. The released definitions preserve exact spelling and hyphenation.

Probability expressions use `the event was {item}`. The items, in order, are impossible, extremely unlikely, very unlikely, unlikely, improbable, possible, somewhat likely, likely, very likely, almost certain, and certain. Their assigned coordinates are 0.01, 0.06, 0.14, 0.24, 0.34, 0.50, 0.64, 0.76, 0.86, 0.94, and 0.99. These define the benchmark's spacing rather than empirical probability calibration. Core target arrays match these definitions up to float32 serialization. The complete-condition figure and numerical tables list all 48 checkpoint names.

A.2 Inputs, layers, and source boundaries

Source and query item IDs are disjoint within each condition. The fitting payload contains source targets and source/query features; query targets reside in a separate evaluation file. The full source is sorted stably by target, and the supplied side identifies the extreme source endpoint. The core fitting sets and query counts are given in Table 1. Evaluation on the opposite side fits a separate predictor, and the two sides can reuse observations in different roles.

The stored-layer convention is $\mathrm{round}(0.75(L-1))$ for a hidden-state array with L entries. For a compact two-dimensional cache, the selected layer is already extracted. This distinction prevents an embedding-state offset from silently changing the layer. Global preprocessing casts the selected features to float32 before randomized PCA with seed zero. PCA has at most $\min(32, n_{\mathrm{fit}} - 1, D)$ components. Source `fit_transform` statistics define coordinate means and scales, and subsequent predictions follow the saved transform path. The local branch uses float64 features and full-coordinate source standardization.

The pseudo-validation set contains the source boundary's $\max\{2, \min(\lceil 0.25 n_S \rceil, n_S - 6)\}$ items. Preprocessing and fitting use only the remaining inner source. Pseudo-validation labels score candidates; final fitting then uses all source labels. The selected configuration is the minimum-loss candidate, with serialized-configuration ordering for ties.

Table 4: Global source candidate grid. Requested dimensions may coincide after rank capping. There are 4,800 nominal settings per condition.

Parameter	Values
Requested dimension	4, 8, 12, 16, 24, 32, 48, 64, 96, 128
Ridge penalty	$10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000$
Scalar continuation	Score-affine, score-piecewise, value-piecewise, distance-affine
Endpoint shrink	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.5, 2, 3, 4, 6

A.3 Scalar continuation and calibration

Let $x(h)$ be the PCA–ridge score. The affine map fits source y on x; the piecewise map interpolates unique sorted source scores, averaging target values at tied scores. Beyond its score support, it uses a least-squares line on the extreme $\max(2, \min(n, \max(3, \lfloor n/4 \rfloor)))$ unique scores. Score-reflection maps first reflect scores on the source side of the endpoint score, with orientation determined by source score–target correlation, and then reflect mapped values to the supplied target side. Value-reflection maps apply the latter step directly. The distance-affine map uses the absolute affine slope and absolute endpoint-score displacement. A constant-score fallback uses the source mean before outward reflection.

Endpoint shrink multiplies displacement from y_e. Equation (1) then fits a nonnegative scale to absolute source displacement. For a fixed upstream score and positive shrink s, its fitted scale

becomes c_1/s when the denominator floor is inactive, leaving the final prediction unchanged. Shrink can still change which upstream configuration the upstream-loss selector chooses. The upstream and final-loss objectives are therefore kept distinct. The complete recorded reselection evaluates $384 \times 4{,}800 = 1{,}843{,}200$ candidates and retains both winners per condition.

A.4 Local relation and direct alternatives

The radial fit in Eq. (3) has an unpenalized intercept and ridge penalty 10^{-3}. The tail size uses Python's round convention, with at least three source items. Radius is standardized on the fitting subset. Constant coordinate scales use the floors stated in the method. Whole-source radius correction changes the subset to S and refits its radial scale, line, and source residual trend; it retains the same global reader, multiplier 3, and gain 0.05.

The direct alternatives share the full-coordinate source standardization. Boundary-feature and all-source-feature ridge regress on these coordinates; boundary-radius and all-source-radius ridge regress on their scalar radius. Each family chooses the eight penalties in the global grid and either direct output or $y_e + d|\hat{y} - y_e|$, a unit-multiplier endpoint reflection. Their validation split is the same source-pseudo boundary. Within-family ties use penalty then output-map name; between-family ties use method name. Thus the source-selected simple reader chooses among 64 candidates. The joint score–radius correction is a separate fixed-penalty diagnostic, not this family selector.

A.5 Correspondence controls

For $k = 3, 5, 6$ tail observations, full-tail derangement enumerates 2, 44, or 265 assignments. Endpoint-preserving derangement fixes the endpoint and enumerates 1, 9, or 44 assignments among the other $k - 1$ observations. Each control refits the local line and residual trend. The source label multiset, endpoint reference, query set, and global branch stay fixed. Losses are computed per assignment and averaged afterward.

Endpoint-indicator correction replaces the radial feature by endpoint membership. Endpoint-deleted fitting removes the endpoint from the regression rows while keeping it as the distance reference. The latter has heavy-tailed errors: one probability/upper condition has nMAE near 97.13. Its mean therefore does not imply that deletion harms every condition. Projecting predictions to the supplied side is recorded as a separate output operation in the complete numerical table.

A.6 Residual identities

Write $\tilde{q}_i = q_i - \bar{q}$ and $r_i = L_i - q_i$. With the variance floor inactive,

$$\gamma = \frac{\sum_i \tilde{q}_i (r_i - \bar{r})}{\sum_i \tilde{q}_i^2}, \qquad \sum_i \tilde{q}_i (C_i - \bar{C}) = 0, \qquad \bar{C} = \bar{r}.$$

Consequently the unpenalized affine fit of p on q has slope one and intercept $0.05\bar{r}$. A source-output match tests this offset explanation; it receives the corrected source outputs and is not an independently trained global-only predictor. Orthogonality to q does not imply independence from nonlinear functions of q or orthogonality on queries.

For constant $L(h) = c$, the same algebra gives $p(h) = q(h) + 0.05(c - \bar{q})$. On a region with fixed reflection sign, write $L(h) = a_\rho \rho(h) + b_\rho$. Then

$$p(h) = [1 - \beta(1 + \gamma)] q(h) + \beta a_\rho \rho(h) + \beta(b_\rho + \gamma \bar{q}).$$

The coefficient 0.05 alone does not bound this correction or guarantee that detrended outputs remain outside the source interval. The no-detrending and mean-removal comparisons below test the specific construction directly.

A.7 Fixed-prediction error accounting and gain sensitivity

Equation (7) follows by separating movement toward and away from a nonzero residual, and then separating moves that cross the target. For all queries, including initially exact predictions, write

$$A = \frac{\sum_j \operatorname{sign}(e_j)\delta_j}{n_Q \sigma_S}, \quad O = \frac{2\sum_j (|\delta_j| - |e_j|)_+ \mathbf{1}\{e_j \delta_j > 0\}}{n_Q \sigma_S}, \quad Z = \frac{\sum_j |\delta_j| \mathbf{1}\{e_j = 0\}}{n_Q \sigma_S}.$$

Then $\mathrm{nMAE}(q) - \mathrm{nMAE}(q+\delta) = A - O - Z$. The analysis checks this identity for each of 384 conditions and both starting readers, with maximum discrepancy 7.07×10^{-17}. Each reader has 2,352 held-query appearances across conditions. Table 3 averages conditions equally; the percentages moving toward the target use the same weighting.

The following sensitivity holds source fits and residual trends fixed and evaluates $q+\beta C$. It describes existing held predictions and does not select a new gain. In particular, the low error at a larger gain for the global reader is not substituted for the source-selected experiments or the frozen headline result.

Table 5: Descriptive gain sensitivity with every fitted component held fixed.

Gain	PCA global plus correction	Simple reader plus correction
0	0.714111	0.531066
0.005	0.705406	0.528667
0.0125	0.692500	0.525288
0.025	0.671641	0.520223
0.05	0.632372	0.512280
0.075	0.595538	0.507312
0.1	0.562658	0.505645
0.2	0.484180	0.524956

B Complete Core Results

The following table retains both global-selection policies and the main correction variants. Its means are recomputed from individual predictions. No entry substitutes a direct predictor for a correction component with the same feature name.

Table 6: Core nMAE by global-selection policy. Both policies evaluate the same 384 conditions.

Method	Upstream loss	Final loss
Global	0.718854	0.714111
Correction	0.634525	0.632372
Endpoint-preserving control	0.648092	0.645342
Full-tail control	0.729237	0.724073
Endpoint indicator	0.678703	0.676309
Endpoint-deleted fit	0.918012	0.915613
All-source radius correction	0.521983	0.528889
Boundary-feature correction	0.709653	0.703714
All-source-feature correction	0.611415	0.613417
Joint score–radius correction	0.610986	0.615748
Source-output offset	0.687425	0.684200
Local reading alone (multiplier 3)	1.894329	1.894329
No detrending	0.622357	0.621606
Mean and slope removed	0.664695	0.660441
Zero-slope relation	0.711441	0.706829
Direct score–radius ridge	0.675894	0.673533
Source scalar piecewise	0.739592	0.740610

Table 7: Task means under final-loss selection; both directions and all 48 checkpoints.

Task	Global	Correction	Difference	Improved
Number	0.651846	0.604333	-0.047513	93/96
Ordinals	0.759443	0.651326	-0.108117	89/96
Percentages	0.674903	0.588875	-0.086028	91/96
Probability	0.770254	0.684953	-0.085301	85/96

Table 8: Direct readers and the unchanged correction. Each family uses 16 source candidates; family selection uses 64.

Reader	nMAE	Change	Rank	Unit
All-source features	0.668144	0.629610	0.314540	0.812295
All-source features + correction	0.578140	0.596175	0.448220	0.745180
All-source radius	0.731237	0.720663	0.791673	0.986643
All-source radius + correction	0.754094	0.716448	0.791673	0.978968
Boundary features	0.745784	0.660477	0.270970	0.856260
Boundary features + correction	0.617091	0.620213	0.483699	0.787535
Boundary radius	0.534938	0.604816	0.786465	0.772441
Boundary radius + correction	0.521423	0.604371	0.786465	0.771083
Selected simple	0.531066	0.616883	0.673047	0.781914
Selected simple + correction	0.512280	0.610204	0.703361	0.763468

The three numerical tasks alone have the results reported in Table 9. They use the same source normalization and equal condition weighting as the full panel; excluding probability codes is a sensitivity analysis.

Table 9: Numerical-only sensitivity: number words, ordinals, and percentages (288 conditions).

Policy	Global	Correction	Reduction
Upstream-loss policy	0.702376	0.619339	11.82%
Final-loss policy	0.695397	0.614844	11.58%

C Uncertainty and Comparison Units

For each contrast, we form a checkpoint-by-task matrix after averaging the two directions. Crossed resampling independently draws checkpoint IDs and task IDs with replacement, keeping their Cartesian-product entries. Task-fixed resampling draws checkpoints while retaining the four tasks. The analysis uses a fixed recorded seed and 20,000 draws. Percentile intervals are descriptive finite-panel sensitivity summaries: checkpoint ancestry and the reuse of four task lists are not removed by resampling. The three focal core contrasts additionally use 98.333% marginal intervals (Bonferroni for three comparisons). Later exploratory contrasts are displayed separately.

Table 10: Paired nMAE differences and descriptive 95% intervals. The first three crossed 98.333% intervals are respectively $[-0.109492, -0.053276]$, $[0.009318, 0.206845]$, and $[-0.027617, -0.003314]$. The last two comparisons are exploratory.

Contrast	Mean	Crossed 95%	Task-fixed 95%
Correction minus global	-0.0817	[-0.1042, -0.0581]	[-0.0889, -0.0740]
Correction minus simple	+0.1013	[+0.0235, +0.1874]	[+0.0826, +0.1190]
Correction minus endpoint control	-0.0130	[-0.0247, -0.0042]	[-0.0160, -0.0102]
Corrected simple minus simple	-0.0188	[-0.0312, -0.0031]	[-0.0250, -0.0133]
Corrected simple minus its control	+0.0007	[-0.0026, +0.0047]	[-0.0010, +0.0022]

For held values y_j and predictions p_j, the all-pairs change error is

$$\frac{1}{\binom{n_Q}{2}\sigma_S}\sum_{j<k}|(p_k-p_j)-(y_k-y_j)|, \qquad \sigma_S=\max\{\mathrm{sd}_S(y),10^{-8}\}.$$

Unit error uses $a=\sum_j(y_j-\bar{y})(p_j-\bar{p})/\sum_j(y_j-\bar{y})^2$ and reports $|a-1|$. Spearman uses average ranks for ties. A constant prediction has undefined rank correlation; the evaluator assigns zero by convention, and that convention is preserved in these aggregate tables. Outside-source nMAE equals nMAE for these range holdouts and is not treated as another independent outcome. Neither the number of item pairs nor the number of bootstrap draws increases the number of task identities.

D Adaptive Support, Gain, and Layer Experiments

The adaptive studies reuse four explicitly separated cohorts: 384 original-core conditions, 16 new-checkpoint core conditions, eight notation-stress conditions, and a separate 12-condition entity cohort. The latter uses a different feature extraction from the two-model geometry experiment in Section 4.7; their entity means are not pooled.

Tail/displacement selection crosses fractions $\{0.2, 0.4, 0.6, 1.0\}$ with multipliers $\{1, 2, 3\}$, keeping layer 0.75 and gain 0.05 fixed. The joint arm additionally selects gain from $\{0, 0.01, 0.025, 0.05, 0.075, 0.1\}$. Two source boundary holdouts, fractions 0.20 and 0.35, refit the full global and local pipelines on the inner source and score the complete final prediction. Ties favor zero gain. Held labels are opened after prediction files are saved. Small source sets can collapse different fractions to the same minimum three observations.

On the original core, the joint selector chooses the full-source tail in 215/384 conditions and gain 0.1, the grid maximum, in 293/384. The result supports broad fitting support but remains sensitive to the chosen candidate class. Separate layer studies use the actual available layers: 0.5/0.75/1.0 for core and notation cohorts, and 0.25/0.5/0.75 for the separate entity cohort. The full joint arm has 216 layer/tail/multiplier/gain settings per condition. Its selected global reader differs from the fixed-layer baseline and is reported beside it.

Table 11: Adaptive tail and displacement, with fixed layer 0.75. Joint selection also chooses gain. The entity cohort here differs from the matched geometry study.

Cohort	Fixed rule	Tail/multiplier	With gain
Original core	0.632372	0.562354	0.541939
New-checkpoint core	0.747152	0.681739	0.615449
Separate entity cohort	1.405590	1.395803	1.390422
Notation stress	0.537236	0.527277	0.527121

Table 12: Joint layer/tail/multiplier/gain extension. Each selected correction is paired with its own selected global reader.

Cohort	Fixed global	Fixed rule	Selected global	Selected rule
Original core	0.714111	0.632372	0.718952	0.554528
New-checkpoint core	0.808885	0.747152	0.760946	0.663006
Separate entity cohort	1.420337	1.405590	1.314893	1.302071
Notation stress	0.530745	0.537236	0.487933	0.507428

Choosing settings on source does not guarantee a better held reading. For example, joint layer selection on notation stress yields corrected nMAE 0.507428 against its own selected global baseline 0.487933. Local-penalty selection over seven penalties gives 0.637598 on the original core versus 0.632372 for the fixed penalty. Complete prediction records for the sweeps, continuous-gain diagnostic, automatic-direction and source-gate diagnostics, and fixed-range source subsamples remain separate numerical analyses in the evidence archive. They do not replace the frozen rule's primary estimate.

E New Extractions and Notation Stress

The two new core checkpoints are `EleutherAI/gpt-neo-125m`, and `EleutherAI/pythia-160m-deduped`. The extraction records preserve full revision IDs and weight, tokenizer, and configuration hashes, token IDs, and carrier positions. An unsuccessful GPU-memory allocation was followed by a documented CPU float32 run. Carrier-token means define the feature vectors. This is a new extraction on existing task definitions, not a new sample of task identities.

The separate Granite-4.0-H-350M-base execution uses a fixed model snapshot, bfloat16 inference, and float32 features with carrier-span mean pooling. Its 70 stored rows have 33 layers and hidden dimension 768. The runner uses Transformers 4.56.0 and the slow GPT-2 tokenizer; complete software and checkpoint identifiers accompany the extraction. It supplies eight conditions under upstream selection, giving global/corrected means 0.720770/0.652587. Independent recalculation from its saved predictions confirms these means; this run does not include a new endpoint-preserving control or a final-loss selector. It remains separate from the two-model fresh cohort and from the original 48-checkpoint panel.

The notation panel contains Roman numerals I–CXXVIII, with targets 1–128, and duration strings in HH:MM:SS format at 37-second increments, with targets 37–4736 seconds. Each model has 102 source and 26 query values per side. There are eight distinct model–notation–direction conditions and two selector arms per condition. All 512 text extractions and all conditions are retained.

Table 13: Separate new-checkpoint and notation panels. Positive percentages denote lower error. The same source-normalized metric is used, but tasks and split proportions differ.

Panel	Policy	Global	Correction	Reduction
New checkpoints / core	Upstream	0.817935	0.755158	+7.68%
New checkpoints / core	Final	0.808885	0.747152	+7.63%
New checkpoints / notation	Upstream	0.423339	0.420697	+0.62%
New checkpoints / notation	Final	0.530745	0.537236	-1.22%

The standalone comparisons on the notation panel were added after seeing the fixed correction's results, using the same 16 settings per family as on the core. Their mean nMAEs are 0.300954 for all-source features, 0.975161 for all-source radius, 0.222396 for boundary features, and 0.254488 for boundary radius. Source/held proportions and value spans differ from the short core, so these panels' absolute nMAEs should not be ranked as if they measured identical tasks.

F Geometry Interventions

F.1 Matched representations and targets

The geometry study uses Qwen3.5-4B and Qwen2.5-0.5B. Each of six entity properties contains 512 records, and the core lists contain 70 records in total. Features are newly extracted with bfloat16 model inference at transformer-depth fraction 0.75; pooled vectors are serialized as float32. Carrier-span means cast hidden states to float32 before averaging. Last-token pooling uses the last non-padding token. Batch size is 16, and extraction does not truncate inputs. The retained receipts record token positions, feature hashes, and local weight-file fingerprints.

Within an entity family, original, explicit-property, long-context, and numeric prompts share item identities, targets, and splits. Numeric prompts supply the recorded target value intentionally. Long prompts alter tokenization, position, and context as well as length. The target coordinates preserve the benchmark transformations, including logarithmic population and area targets where specified; changes in those coordinates are not reported as raw physical increments. Entities can occur in more than one attribute family.

The main 224 conditions comprise two models, two pooling choices, two sides, six attributes with four prompt forms, and the four core lists. Additional studies include 120 source-size conditions, 144 constructed spacing/range conditions, and 64 source-readout conditions. These counts describe overlapping evaluations.

F.2 Distance interventions and source selection

Euclidean distance uses full source-standardized coordinates. Cosine distance, PCA, random projection, source-supervised PLS, and a source nearest-neighbor graph provide alternatives. Projection dimension is four, reduced to the feasible source rank; random and supervised projections use matched dimensionality. Graph construction uses symmetric union nearest neighbors, with $k = 10$ capped by source size and increased when needed for connectivity. A query connects to the fixed

source graph, without query–query edges. The graph's direct-regression control uses the original standardized features, not a separately learned geodesic embedding.

Distance interventions share the global prediction, direction, labels, split, multiplier 3, and gain 0.05, then refit the local relation and detrending. The global policy here retains the upstream-loss selection objective. This keeps the metric comparison separate from changing global selection. Direct readouts in this study use fixed ridge penalty 0.001; they are not the 64-candidate core reader family.

Gain-only selection holds Euclidean distance and tail 0.4 fixed and chooses $\beta \in \{0, 0.01, 0.025, 0.05, 0.1\}$. Metric/tail/gain selection crosses six distances, three tail fractions $\{0.1, 0.2, 0.4\}$, and these five gains. Source-pseudo holdouts use fractions 0.20 and 0.35, removing duplicate splits on small lists. Every learned transform and prediction stage is refitted inside each fold. The geometry gate uses nonendpoint source-tail correlation at least 0.5 to allow the selected gain and otherwise sets gain to zero. This is a fixed diagnostic threshold.

Table 14: Geometry study: original prompts and carrier-span mean pooling, averaged over both models and directions. Fixed metric rows use tail 0.4 and gain 0.05.

Reading	Core nMAE	Entity nMAE
Global	0.749923	1.356418
Euclidean	0.699059	1.352185
Cosine	0.727217	1.350657
PCA4	0.730639	1.352808
Random4	0.701592	1.353066
PLS4	0.722815	1.351349
Graph	0.692952	1.352613
Selected gain	0.705398	1.352384
Selected metric/tail/gain	0.689733	1.336947
Geometry-gated gain	0.745221	1.356418

Table 15: Matched entity prompts and pooling. Numeric prompts intentionally contain the target value.

Prompt	Pooling	Global	Euclidean	Selected
Original	mean	1.356418	1.352185	1.336947
Original	last	1.397747	1.382827	1.356831
Explicit	mean	1.362155	1.358297	1.337678
Explicit	last	1.324140	1.316879	1.298703
Long	mean	1.328935	1.331990	1.310393
Long	last	1.361436	1.359726	1.342937
Numeric	mean	1.378509	1.363422	1.338985
Numeric	last	1.426037	1.412998	1.397401

F.3 Information diagnostics and alignment

Random five-fold source readouts are a retrospective diagnostic of information accessible inside the source range. Each fold refits normalization, PCA, PLS, and radius from its training observations. They do not evaluate the held target range. Source distance–attribute association excludes the endpoint; tail refit R^2 also removes and refits without the endpoint. Undefined small-tail correlations remain missing. The field `radius_cv` in the data denotes radius coefficient of variation, not cross-validation performance.

The geometry intervals use 20,000 family-cluster bootstrap draws, retaining models and directions within a family. There are four core or six entity families. These intervals measure sensitivity to the observed families; overlapping entities and small family counts constrain population interpretation. The selected metric/tail/gain contrast is relative to the fixed Euclidean correction, not automatically relative to the global reader.

Table 16: Random five-fold source readout R^2. The three columns have different input conditions; these are source interpolation diagnostics.

Readout	Core	Entity	Number input
Constant	-0.1810	-0.0029	-0.0029
Five nearest neighbors	0.2465	0.3261	0.5345
Whole-source radius	-0.2650	0.0614	0.0100
Full-feature ridge	0.5429	0.1934	0.8736
PCA4 ridge	0.4742	0.2829	0.2245
PLS4 ridge	0.5427	0.4798	0.6613
Random4 ridge	-29.5603	0.0562	0.0972

Table 17: Endpoint-excluded alignment on original entity inputs. Tail R^2 is refitted without the endpoint. Radius CV is coefficient of variation.

Metric	Source rank	Held rank	Tail R^2	Radius CV
euclidean	0.1963	0.0815	0.0198	0.1190
cosine	0.2046	-0.0056	0.0469	0.1423
pca4	0.1874	-0.0599	0.0253	0.4028
random4	0.0367	0.0249	0.0141	0.3381
pls4	0.4555	-0.0175	0.0524	0.4193
graph	0.1713	0.0291	0.0222	0.2699

F.4 Sample size, range, and spacing

The source-size experiment uses 32, 64, 128, 256, and 410 rank-spaced source observations while holding the full-source endpoints and query items fixed. Its nMAE denominator is the original full-source standard deviation, so a smaller training set does not change the scale. Each size uses one deterministic subset, not a new independent dataset.

Constructed numerical inputs compare empirical quantiles, uniform spacing, quadratic gaps, and the central half of the range at 128 items, plus uniform sets with 32 and 512 items. Values stay in the benchmark target coordinates. These are number-reading interventions, not newly fabricated entity properties. Their condition-specific source standard deviations mean that narrower-range nMAE must be distinguished from absolute error.

Table 18: Source-size intervention on entities, with fixed endpoints, queries, and full-source normalization.

Source size	Global	Euclidean	PLS4	Graph
32	1.5268	1.5088	1.5071	1.5083
64	1.3244	1.3123	1.3108	1.3125
128	1.4075	1.3977	1.3946	1.3978
256	1.3637	1.3583	1.3570	1.3585
410	1.3564	1.3522	1.3513	1.3526

Table 19: Constructed number-input range/spacing conditions. Normalization uses each condition's source SD.

Spacing	Items	Global	Euclidean	PLS4	Graph
empirical	128	1.5424	1.5276	1.5292	1.5294
half range	128	0.9642	0.9871	0.9851	0.9875
quadratic	128	0.7736	0.7916	0.7874	0.7926
uniform	32	0.8594	0.8961	0.8882	0.8820
uniform	128	0.9367	0.9591	0.9566	0.9565
uniform	512	0.9567	0.9777	0.9765	0.9781

G Reproduction and Numerical Checks

The supplementary materials provide code, source labels, evaluation targets, and prediction records. The compact archive supports prediction-level replay; high-dimensional activation arrays and exhaustive candidate-score caches are omitted to meet the upload limit. The manuscript's table and figure data are exported directly from prediction-level analyses. The core evaluator checks IDs, source/query separation, range boundaries, correction formulas, endpoint-preserving permutations, selected source candidates, and saved prediction identities. Re-execution passes 17,956 checks. A separate SVD implementation verifies 24,576 simple-reader candidate losses and 1,536 final fits, with maximum query prediction discrepancy 6.31×10^{-10}. These counts describe arithmetic checks, not experimental sample sizes.

The geometry audit reconstructs 4,032 radial fits and corrections across all 224 main conditions using separate closed-form arithmetic; maximum absolute discrepancy is 1.78×10^{-15}. Complete fitting is replayed for four fixed real-data conditions, reproducing stored predictions exactly. Euclidean, cosine, PCA, random, PLS, and graph query distances agree under singleton and reversed-order evaluation within 2.14×10^{-13} in those checks. The fresh-core and notation prediction evaluators pass 1,056 and 528 checks, respectively. These are numerical reconstructions of saved predictions, distinct from rerunning every language-model extraction.

Numerical portability. The primary computation uses Python 3.13.12, NumPy 2.4.4, SciPy 1.17.1, and scikit-learn 1.9.0 on Linux aarch64. A second execution uses Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0 on Linux x86_64. Re-evaluation passes all 17,956 prediction checks. Saved-configuration refits change the final-loss global and corrected mean nMAEs by less than 8×10^{-7}. Full source reselection changes 585 of 768 nominal configuration strings. The candidate space includes rank-capped dimensions, calibrated shrink factors, and nearly tied source objectives; configuration identity and prediction agreement are therefore distinct checks. The supplementary materials include the two environments and the complete numerical comparison.

The mathematical predictor is pointwise, while floating-point execution can depend on matrix shape. Recorded float32 PCA singleton/batch drift is at most approximately 0.000168 source standard deviations on the core. In the duration stress test it can be amplified by scalar continuation to 16.27 seconds, or 0.01494 source standard deviations. Singleton evaluation changes mean stress nMAE to 0.420689 and 0.537309 for the two policies, preserving the qualitative result. This is numerical variability, rather than fitting a coordinate from query ranks.

H EVERY CORE CONDITION

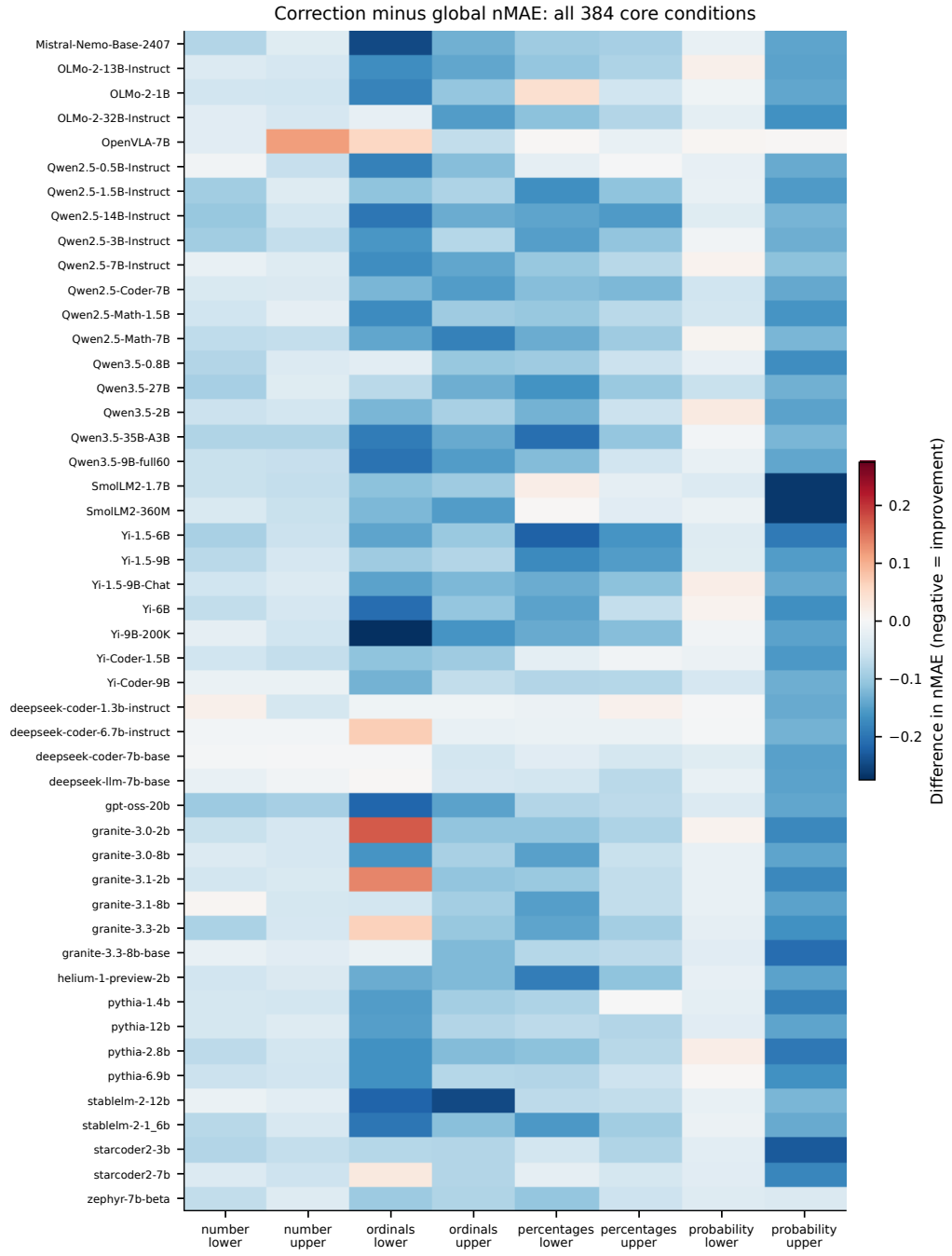

Figure 6: All 48 checkpoints, four tasks, and both directions under final-loss selection. Entries are paired corrected-minus-global nMAE. Colors share a single scale, and all 384 conditions are retained. Exact values are available in the supplementary CSV.