BEYOND CHARACTER INDICES: EXPLICIT GEOMETRY FOR PROPORTIONAL TEXT ART

Anonymous authors
Paper under double-blind review

ABSTRACT

A spatial quantity can be fully determined by a model's inputs yet remain useful to supply explicitly. We test this distinction in proportional-font text art, where horizontal position is the cumulative sum of known glyph widths. Pixel-preserving whitespace rewrites change character indices while holding the rendered artwork fixed. Paired Transformers share characters, true widths, target inputs and complete source access, differing only in an extra source scalar: matched index or cumulative position. Canonicalizing every source run in the specified whitespace family during training and evaluation makes each model exactly invariant to those rewrites. Cumulative positions still improve conditional nonspace prediction by 0.0365 bits/target. Three additional paired initializations at a fixed endpoint retain a 0.0410-bit gain. Width-only, calibrated row-scale and pseudo cumulative controls also favor cumulative positions. Frozen models retain a gain on 19 newly acquired, screened event artworks, and larger predictors retain gains with both normalized and raw/augmented inputs. In this controlled natural setting, explicit cumulative position remains useful after the specified encoding invariance is enforced.

1 INTRODUCTION

A structured quantity can be determined by a model's inputs without being equally easy for that model to use (Xu et al., 2020b). Spatial position in proportional text provides a concrete example: a character's horizontal coordinate is the sum of the preceding character widths. Giving a predictor every local width supplies the ingredients of that computation; giving it the sum specifies their spatial relation. Does supplying the sum improve prediction even when redundant encodings of the same rendering have already been removed? Proportional-font text art lets us control local-width access and encoding redundancy separately.

Our setting is Shift_JIS art, the multiline text-art tradition known as AA in Japanese online communities, with a repertoire extending beyond ASCII (Carlsson, 2017; Mona Font Project, 2005). In the MS PGothic proportional-font convention, characters and whitespace jointly form a spatial composition. A glyph can contribute a short stroke to a contour that continues on another row; spaces place that stroke relative to the surrounding marks. Readers perceive the arrangement across rows as an artwork. Figure 1 places a prediction input within its complete natural artwork, so that the selected rows can be read as parts of that composition.

The relevant horizontal relation is measured in rendered distance. Character indices count symbols, but symbols with different advances occupy different amounts of that distance. Consequently, characters at the same ordinal column on adjacent rows need not align, and the cumulative widths of their prefixes determine their horizontal positions. Local glyph identity and cross-row alignment therefore provide distinct cues to a character's role in the image.

This natural setting also supplies an exact intervention on the encoding. Five ideographic spaces and eleven ordinary spaces both advance 55 pixels under the fixed rendering rule. Exchanging them in the illustrated source row changes subsequent indices while preserving the complete rendered pixels. The known rendering rule lets us vary a string representation while holding its depicted structure fixed. This separates the value of explicitly supplied position from a change in the visual content being predicted.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

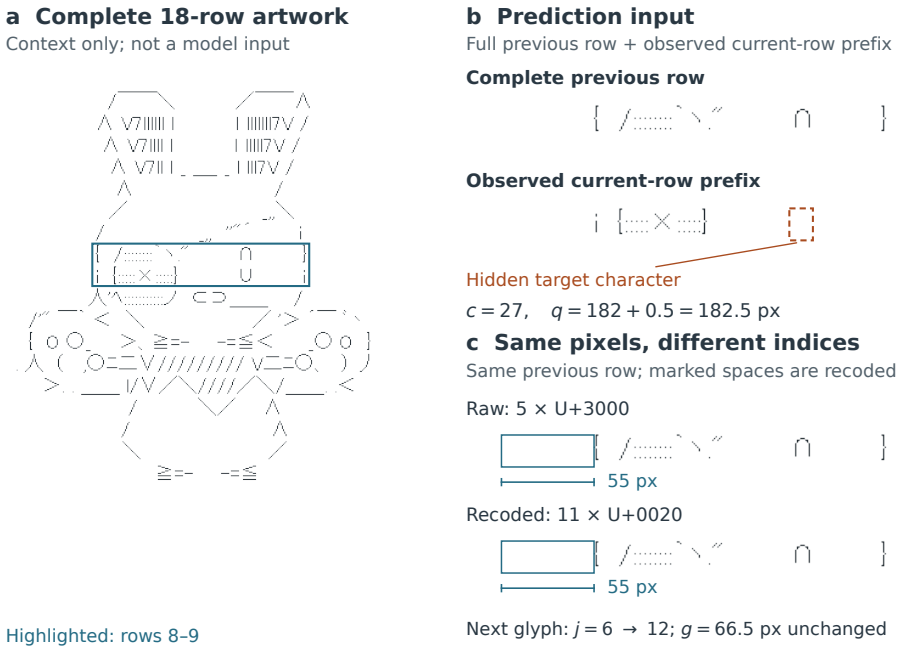

Figure 1: Character indices and rendered positions can vary independently. **(a)** The complete 18-row artwork provides context; outlined rows 8–9 form the example. **(b)** The model receives the complete previous row and observed current-row prefix. That prefix determines $c = 27$ and $q = 182.5$ pixels. The fixed-width dashed empty box marks the hidden target without depicting its glyph or width. **(c)** The same previous row after replacing the marked five ideographic spaces by eleven ordinary spaces, retaining the first ideographic space. Both marked spans occupy 55 pixels under MS PGothic at 16 pixels; complete-row and artwork pixels are identical. The first visible source character changes index $j = 6 \rightarrow 12$ while its position $g = 66.5$ pixels stays fixed. The current-row prefix, target and query are unchanged.

We predict a nonspace character using the completed previous row and the observed current-row history. Our *width-known* comparison gives W-index and W-scan the same characters, true local widths, target geometry and unrestricted query access to source positions. Their active projection interfaces have the same architecture and parameter count. The intervention is whether an additional matched scalar supplies a character index or an explicitly accumulated coordinate (Figure 2).

The experiments establish the prediction benefit, test which position information supports it, and examine its persistence across materials and model settings:

- **A matched test beyond exact encoding invariance.** Shared true widths and full source access, together with source canonicalization, let us test the value of an explicit relation with local measurements held fixed and the specified encoding redundancy removed. The original and fixed-endpoint additional stages retain gains of 0.0365 and 0.0410 bits/target (Section 3).
- **Value of the true cumulative metric.** Cumulative positions improve over width-only, matched-index and calibrated row-scale inputs, and over a fixed pseudo cumulative metric. Index itself improves on width-only; the Index–Scan gain also persists with residual dropout (Section 3).
- **Persistence across materials and model settings.** Frozen models retain a gain on 19 screened event artworks. The benefit also persists in a larger regularized normalized model and in separate raw/augmented causal predictors (Sections 3–4).

The resulting test separates access to local measurements, invariance to specified encodings and the predictive value of an explicit spatial relation.

2 A MATCHED TEST WITH KNOWN LOCAL MEASUREMENTS

2.1 PREDICTION TASK AND EXPLICIT POSITIONS

For a target character $x_{r,c}$ at zero-based column c, define

$$q(r,c) = \sum_{j<c} w(x_{r,j}) + 0.5, \qquad g_j = \sum_{t<j} w(x_{r-1,t}) + 0.5. \tag{1}$$

The fixed MS PGothic 16px advance table supplies w. The target query q uses only the observed prefix, not the target character's width or a future suffix. Both arms receive the complete previous row, source-index sinusoidal encodings, up to 128 preceding current-row characters, and the same target summary (c, q). The full known prefix determines q even when it exceeds the history window. Target-side accumulation is thus already supplied in both arms; the intervention concerns source position.

Eligible targets are nonspace characters below the first row, in rows containing at least two ink characters. Unknown targets and empty sources use shared smoothing and masking rules. Appendix G specifies eligibility, the output distribution and preprocessing.

Each paired evaluation fixes the previous row, observed history, query and target, without generated errors changing later inputs. In Figure 1, the target completes the lower outline of the artwork's right eye; the source row contains its upper outline.

NLL measures the actual character's probability; accuracy measures the highest-probability choice. Main tables evaluate the complete prescribed population with the weights below.

2.2 MATCHED REPRESENTATIONS AND SOURCE ACCESS

Both width-known models receive the true advance of each source character, looked up before vocabulary mapping, including characters whose token IDs become UNK. Their inputs to an active, bias-free $2 \to d$ projection are

$$\text{W-index: } [\widetilde{w}(x_j), h_j], \qquad \text{W-scan: } [\widetilde{w}(x_j), \widetilde{g}_j]. \tag{2}$$

Tildes denote training-fitted normalization. Positive uncentered widths are summed before normalization. The affine index scalar h_j matches the mean and variance of $\widetilde{g}_j$ over valid training source-position occurrences. Padding is excluded. Cumulative coordinates express the source in the target query's spatial metric; the comparison tests the predictive value of explicitly supplying this computed relation when both arms have its local measurements, the same active projection architecture and equal parameter counts.

Both arms use Transformer encoders (Vaswani et al., 2017). Source and target-history streams share weights within each model; each arm is fitted separately. The query cross-attends to every valid source position without a column or geometric-neighborhood mask. We test both causal and bidirectional source self-attention, while target history always remains causal. These masks change source-state computation while preserving full query access. Figure 2 locates the intervention.

2.3 MATERIAL, TRAINING AND UNCERTAINTY

The material derives from HukuTemp, a versioned archive of the community-maintained Yaruo-thread MLT Collection (Yaruo-thread MLT Collection, 2026). The inherited pools use a fixed v32.0 derivative snapshot (1anon8anon1, 2026); ordinary additions come from archived v32.2 relative to v32.1. Appendix H.1 records their origins and version identities.

The inherited training, development and original evaluation pools contain 638, 654 and 632 records, respectively, with 537,584, 549,663 and 497,645 targets. Near-duplicate component identities are disjoint across these pools. The main ordinary evaluation consists of 2,951 screened HukuTemp v32.2 additions relative to v32.1, containing 2,662,587 targets in 611 source-file blocks. Screening removes detected overlap with bound older collections and usage records. This material has been used in preceding research decisions and remains within the same rendering tradition; it is not a new blind confirmation or an independent-author split (Appendix H). A separate frozen-model evaluation uses newly acquired, screened event artwork (Appendix A.3).

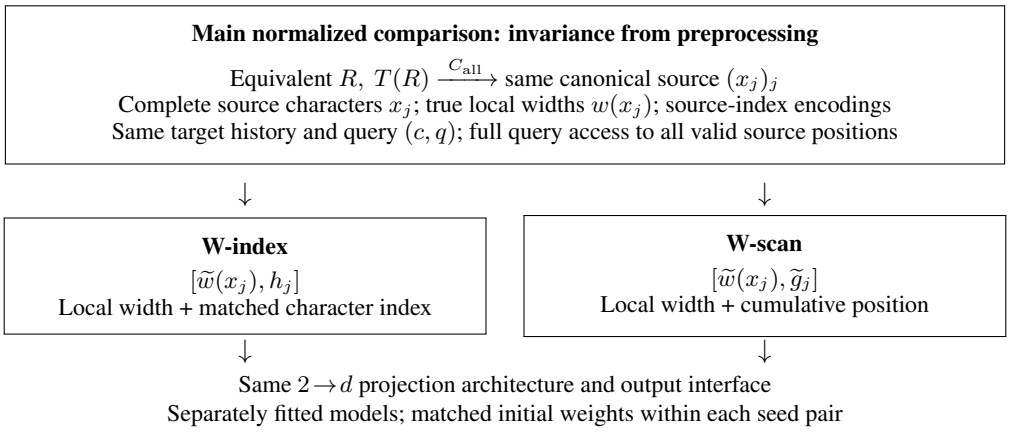

Figure 2: **Preprocessing gives invariance; the scalar interface tests prediction.** In the main normalized comparison, $C_{\text{all}}(R) = C_{\text{all}}(T(R))$ for the specified U+0020/U+3000 rewrite family; separate raw/augmented fits omit this preprocessing. Both arms retain true source widths, including UNK-mapped characters, and full query access. Training-fitted moments match the index and cumulative-position scalars.

Each reported stage averages three paired seeds; normalization has an original stage and a separate fixed-endpoint repeat stage. We average target log scores within records, records within source-file blocks, blocks equally, then seeds equally:

$$\widehat{L}_a = \frac{1}{3B} \sum_{s=1}^{3} \sum_{b=1}^{B} \frac{1}{n_b} \sum_{i \in b} \frac{1}{T_i} \sum_{t=1}^{T_i} -\log_2 p_{a,s}(x_{i,t}). \tag{3}$$

Positive $\Delta = \widehat{L}_{\text{index}} - \widehat{L}_{\text{scan}}$ favors explicit positions. Accuracy uses the same weights. We average losses, not ensemble probabilities. The event evaluation averages within artwork and visible-signature group; the stress population has its own weights (Section 5). These populations are reported separately.

For the ordinary material, bootstrap intervals resample source-file blocks while keeping the three fitted pairs fixed and paired. We also show all seed differences and descriptive Student-t intervals for their mean. These summaries distinguish material uncertainty conditional on selected models from fit variation under a normal-difference assumption that three runs cannot establish. Neither includes pilot-selection or research-adaptation uncertainty. Appendix L records each original endpoint family and its role in the present synthesis.

3 Cumulative positions retain value after whitespace normalization

Does explicit cumulative position still help when equivalent whitespace strings have already been reduced to a single encoding? We test this with C_{all}, which maps every maximal source run of U+0020 and U+3000 spaces to a unique encoding of its total advance. It covers leading, internal and trailing runs and preserves the order and pen origins of nonspace characters. The canonicalization rule and its pixel checks are specified in Appendix A.

Both width-known models are trained and evaluated on these canonical source rows. Histories, targets and target queries remain unchanged; source features and training calibrations are rebuilt from the canonical rows. The models use bidirectional source encoding with two blocks and $d = 64$ (348,811 parameters). Both receive true local widths and use the same training recipe. In the original stage, they share a development-selected epoch.

W-scan improves NLL by **0.0365 bits/target** and accuracy by 0.348 percentage points after both local-width access and the specified encoding redundancy are controlled (Table 1, original stage). The NLL reduction is 1.06%; the corresponding weighted conditional perplexity falls by 2.50%.

Table 1: **Cumulative positions improve prediction in both normalized training stages.** The original three pairs select a common epoch 8; the additional three use that endpoint without reselection. Both stages use the same ordinary material and weights. Material intervals condition on three fitted pairs; fit intervals describe their mean variation.

	Original three pairs	Additional three pairs
Index / Scan NLL	3.4450 / 3.4085	3.4412 / 3.4002
Index / Scan accuracy (%)	51.338 / 51.687	51.293 / 51.706
NLL gain (bits/target)	**0.0365**	**0.0410**
Material 95% interval	[0.0332, 0.0401]	[0.0379, 0.0442]
Paired-fit differences	0.0185 / 0.0459 / 0.0453	0.0489 / 0.0466 / 0.0276
Descriptive fit 95% interval	[−0.0024, 0.0754]	[0.0120, 0.0701]

Because $C_{\mathrm{all}}(R) = C_{\mathrm{all}}(T(R))$, each arm receives identical inputs across the specified rewrites. Both therefore have exact invariance to that family, yet the prediction gap remains.

In the original three pairs, 537 of 611 source-file blocks have positive seed-averaged gains (median 0.0327 bits); Appendix A.5 gives the complete block and position-offset descriptions.

Does the gain persist across new initializations? With the recipe and epoch-8 endpoint fixed before further training, three new paired initializations retain a **0.0410-bit/target gain** and improve accuracy by 0.412 percentage points. All three added pairs favor W-scan, as do all three original pairs (Figure 3). The fixed-endpoint stage is reported separately from the original fits used for endpoint selection (Appendix A.1).

Which source representation provides the gain? Scan improves on Width-only by **0.0601 bits/target** and on Linear by **0.0368 bits/target**, with all three paired differences positive in each comparison. Index also has lower NLL than Width-only in every pair. Thus, under this fixed recipe, Scan's benefit extends beyond both local-width inputs without an extra index scalar and a calibrated coarse row-scale coordinate.

The controls separate four questions. **Width-only** sets the second scalar to zero: is an extra position scalar needed? **Index** supplies a matched ordinal channel; its comparison with Width-only also checks whether that extra index harms prediction. **Linear** spreads total source advance uniformly across indices and matches Scan's training-source moments: does coarse row scale suffice? **Pseudo**, examined next, tests a specified alternative cumulative metric. All retain true local widths. Width-only removes a feature; the other controls retain active position channels. Width-only and Linear use the same three untrained initial states, data orders, recipe and epoch 8 as the additional Index/Scan stage. Table 2 reports the primary Width-only–Scan comparison and secondary Linear–Scan diagnostic on the same canonical targets (Appendix A.2).

Table 2: **Position representations with shared true local widths.** Index and Scan reuse the additional fixed-endpoint stage's three fits; Width-only and Linear add three fits each. NLL is in bits/target and accuracy in percent. Positive contrasts favor Scan. Material intervals condition on three paired fits; fit intervals describe their mean. Linear is a prespecified secondary diagnostic.

Model	Second source scalar	NLL	Accuracy (%)
Index (reference)	Matched index	3.4412	51.293
Scan (reference)	Cumulative position	3.4002	51.706
Width-only (new)	Zero	3.4603	51.059
Linear (new)	Calibrated row scale	3.4370	51.391

NLL contrast	Mean	Material 95% CI	Descriptive fit 95% CI
Width-only − Scan	0.0601	[0.0566, 0.0637]	[0.0514, 0.0689]
Linear − Scan	0.0368	[0.0344, 0.0391]	[0.0057, 0.0678]

Scan also improves NLL by 0.0412 bits/target over Pseudo (Table 3). Pseudo sums fixed positive pseudo-widths, preserving monotonicity and true whitespace advances, with training moments matched to Scan. This tests the specified cumulative metric; the target query remains in the true metric, so accumulation and source–target metric consistency remain intertwined.

Residual dropout. Index/Scan trained with residual dropout 0.1 retain a 0.0242-bit gain, extending the finding beyond the zero-dropout recipe. Both this comparison and Pseudo/Scan retrain both arms at the fixed epoch-8 endpoint; all three paired differences are positive in each family (Appendix B).

Table 3: **Normalized prediction gains under metric and training controls.** Rows are separate six-model bidirectional diagnostics on reused ordinary material. The first two reuse $d64/2$ seed identities at epoch 8; $d128/4$ uses new paired seeds and a common development-selected epoch 6 (Appendix B.3). Individual 95% intervals do not imply joint coverage.

Comparison	Mean	Material 95% CI	Descriptive fit 95% CI
Pseudo − Scan, dropout 0	0.0412	[0.0376, 0.0445]	[0.0136, 0.0688]
Index − Scan, $d64$, dropout 0.1	0.0242	[0.0218, 0.0266]	[0.0131, 0.0353]
Index − Scan, $d128$, dropout 0.1	0.0239	[0.0200, 0.0273]	[−0.0017, 0.0495]

Does the gain persist on new screened material? Frozen models retain a **0.0764-bit/target gain** on newly acquired artwork from the 22nd Kōhaku AA Battle, with all three fitted pairs positive (Table 4). The additional stage's six models, preprocessing, inclusion rules and analysis are fixed before event scoring. Of 49 fixed candidates, screening quarantines 29 with detected relations to prior material and excludes one unsupported by the frozen advance table, leaving 19 before model scoring (Appendix A.3). This extends the fixed-model finding to screened material within the same rendering tradition; populations remain separate.

In the subsequent diagnostic on all 19 works, Scan also outperforms the frozen Width-only and Linear models (Table 4). The Index–Scan gain is positive in 17 of 18 signature groups and remains positive after any one group's removal (Appendix A.4).

Table 4: **Frozen models retain a prediction benefit on screened event artwork.** The 19 works contain 24,748 targets in 18 visible-signature groups. Index/Scan were fixed before event scoring; Width-only/Linear are subsequent diagnostics using already trained models. Group intervals condition on these fits and the screened frame; fit intervals describe three paired differences. Visible signatures do not establish independent authors.

Representation	NLL (bits/target)	Accuracy (%)
Width-only	4.8967	36.183
Index	4.8697	36.097
Linear	4.8525	36.192
Scan	**4.7933**	**36.984**

NLL contrast	Mean	Group 95% CI	Descriptive fit 95% CI
Index − Scan (frozen)	0.0764	[0.0457, 0.1099]	[−0.0445, 0.1972]
Width-only − Scan	0.1034	[0.0461, 0.1786]	[0.0492, 0.1576]
Linear − Scan (secondary)	0.0592	[0.0296, 0.0933]	[−0.0166, 0.1349]

4 LARGER PREDICTORS RETAIN A PREDICTION BENEFIT

A four-block bidirectional $d = 128$ model (1,181,387 parameters) also retains a **0.0239-bit/target gain on fully normalized inputs** (Table 3). Six fits complete eight epochs with residual dropout 0.1; the common six-model development rule selects epoch 6 from 1/2/3/4/6/8. Index/Scan NLLs are 3.4112/3.3872, with accuracy improving by 0.386 percentage points. All three paired gains are positive; the fit interval crosses zero. This crosses normalization with a larger regularized system, changing capacity and training recipe jointly (Appendix B.3).

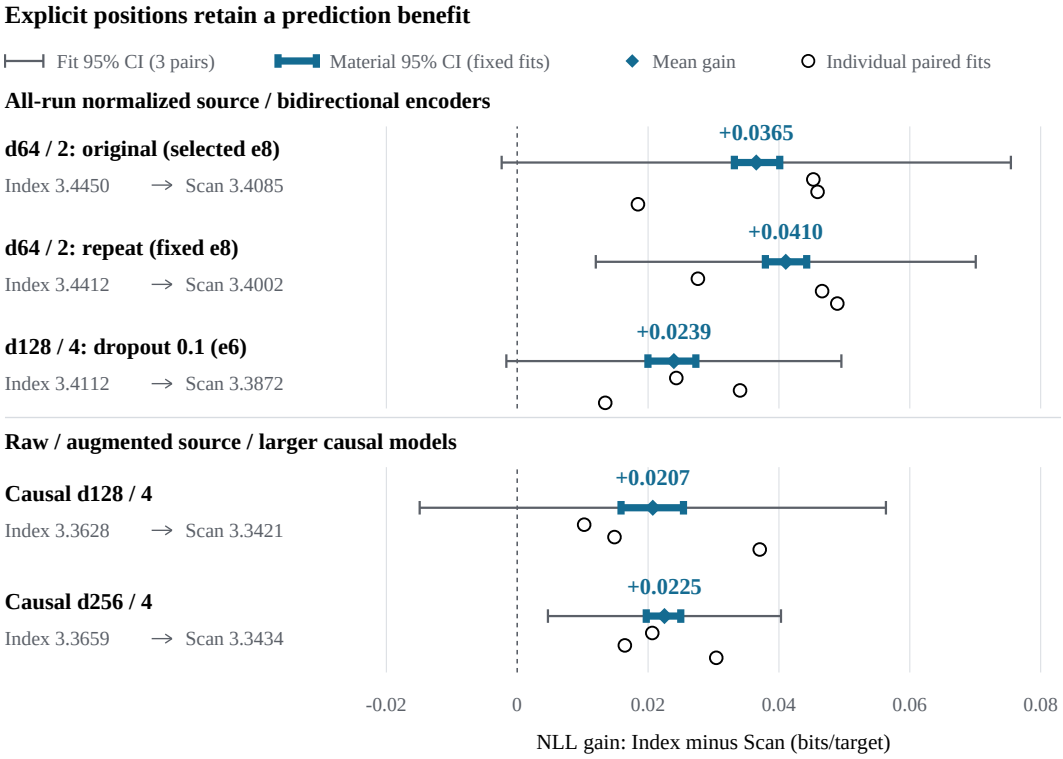

Figure 3: **Prediction gains with normalized inputs and in separate raw/augmented systems.** The upper group separates the two $d64$ stages from the larger regularized normalized model; the lower group shows causal raw/augmented models. Each row reports three pairs and Index-to-Scan NLLs. Diamonds mark means, thick bars conditional block-bootstrap 95% intervals, thin bars descriptive fit t intervals ($df = 2$), and circles individual gains. Vertical offsets separate points only. Neither interval includes selection or research adaptation.

A separate raw/augmented comparison tests four-block causal W models at $d = 128$ and $d = 256$ against the two-block $d = 64$ reference. These experiments share local widths and the same recoding schedule: eligible targets cycle through raw, raw, left-one-step and right-one-step views, replacing exposures rather than adding data.

Both arms receive equal development-only searches over two AdamW (Loshchilov & Hutter, 2019) recipes and five epochs (4/8/12/16/20): a pilot chooses each arm's recipe, followed by three paired formal fits and per-arm endpoint selection. Both $d = 256$ arms select B/epoch 4; $d = 128$ selects B/epoch 4 for index and A/epoch 8 for scan. Appendix C gives the recipes, curves and protocol amendments.

Table 5: **Larger causal models improve absolute prediction while retaining the representation gap.** All rows use shared local widths and the same ordinary evaluation and raw/augmented training schedule. The $d = 64$ reference uses one recipe and a common six-model endpoint; the larger configurations use equal per-arm pilot searches. Cross-row differences therefore do not isolate capacity alone.

Model	Parameters	Index NLL	Scan NLL	NLL gain	Accuracy gain (pp)
$d = 64$, 2 blocks	348,811	3.3930	3.3632	0.0298	0.200
$d = 128$, 4 blocks	1,181,387	3.3628	3.3421	**0.0207**	0.262
$d = 256$, 4 blocks	3,924,555	3.3659	3.3434	**0.0225**	0.228

The $d = 256$ **gain is 0.0225 bits/target** under shared recipe B/epoch 4 selection; every pair favors Scan. The $d = 128$ gain is 0.0207 bits/target under separate selections. Both larger arms improve selected development and ordinary NLL over their small counterparts (Table 5). Their Index NLLs approximately match the small Scan model.

Appendix C retains the development curves, $d = 128$'s recipe/endpoint sensitivity and the reversed single-seed $d = 256$ recipe-A pilot.

Three of four larger-model selections fall at the earliest candidate, epoch 4, and later development losses rise. These gains characterize the selected raw/augmented causal predictors; their search does not include denser early checkpoints or dropout. The normalized $d64$ residual-dropout comparison is reported separately in Table 3.

5 Prediction behavior under unchanged rendering

On uncanonicalized inputs, explicit cumulative positions also reduce average sensitivity to pixel-preserving rewrites, providing complementary behavioral evidence. We measure this on targets with an eligible internal whitespace run on each side of the query. One step exchanges five ideographic spaces for eleven ordinary spaces within a run: both groups occupy 55 pixels, so the rewrite adds six characters while preserving total advance. The raw, L1, R1, L2 and R2 views preserve the target, history, query and whole-row pixels. At each intensity, left and right rewrites match total token count and space composition. Appendix D specifies the run boundaries relative to the query.

The stress population contains 512 components, 1,069 source rows and 6,493 targets selected by input eligibility before reading the new candidates or model outcomes under the recorded-use exclusions. All 25,972 recoded target–view comparisons pass whole-row pixel equality. We weight components equally, rows equally within component and targets equally within row. This population measures behavior on targets with bilateral internal whitespace; it is enriched for rewrite eligibility and does not estimate ordinary-target prevalence. The source-mask and capacity comparisons use these same targets (Appendix D).

Let $J_a(S, 2)$ be the weighted base-2 Jensen–Shannon (JS) divergence between raw and two-step side-S predictions. The original stress study designated

$$K_2 = \{J_{\text{index}}(L, 2) - J_{\text{index}}(R, 2)\} - \{J_{\text{scan}}(L, 2) - J_{\text{scan}}(R, 2)\} \tag{4}$$

as its primary endpoint. Its original endpoint status and interval are retained; here it provides supporting behavioral evidence alongside the ordinary prediction comparisons. A positive value means a larger *absolute* JS reduction before the query. Table 6 gives the corresponding behaviors across configurations; all use the uncanonicalized five views.

Table 6: **Two-step encoding responses on the same stress targets.** Entries average three fitted pairs. JS is in bits; flips are raw-to-L2 argmax changes, not additional errors. The two small-model rows isolate source-mask changes at fixed capacity; the larger causal rows use their own development search.

Source encoder	Before-query JS Index	Before-query JS Scan	K_2	Before-query flips (%) Index → Scan
Causal $d64/2$	0.0248	0.0154	0.0071	18.37 → 14.05
Bidirectional $d64/2$	0.0295	0.0160	0.0089	21.51 → 14.09
Causal $d128/4$	0.0154	0.0097	0.0052	15.27 → 10.95
Causal $d256/4$	0.0215	0.0091	0.0091	18.96 → 12.10

Across configurations, average sensitivity falls on both sides and every pair has positive K_2. The bidirectional small model also retains a 0.0387-bit/target ordinary gain, so both findings occur with full source encoding.

Uniform proportional contraction on both sides can produce positive K_2; the interaction does not identify an internal accumulation algorithm. The complementary log-ratio contrasts have mixed seed signs, and some individual right-side JS differences favor index (Appendices D and C). Those analyses delimit the behavioral finding as lower average encoding sensitivity. Full canonicalization

would merge these views exactly for either model; the learned response measured here concerns uncanonicalized encodings.

6 Related work and scope

Explicit and learned positions. CoordConv demonstrates benefits from explicit coordinate channels in convolutional models (Liu et al., 2018); Contextual Position Encoding constructs positions from learned context-dependent increments (Golovneva et al., 2024). Abacus embeddings provide positions within numbers, while position coupling aligns positions across task-related digits to improve arithmetic length generalization (McLeish et al., 2024; Cho et al., 2024). Our setting makes the incremental value of a fixed cumulative metric directly testable: both models receive its local measurements, specified redundant strings can be identified exactly, and ordinal position can change while the rendered pixels remain fixed.

Computable structure and model behavior. Sinusoidal encodings, rotary embeddings and causal mechanisms offer different routes to ordinal position (Vaswani et al., 2017; Su et al., 2024; Kazemnejad et al., 2023). RASP formalizes Transformer-like counting and prefix operations (Weiss et al., 2021); work on RASP-L relates learned length generalization to program structure and data diversity (Zhou et al., 2024). Gurnee et al. (2025) trace learned accumulation of token lengths in a pretrained model's fixed-width linebreaking. Our comparison tests the predictive value of explicitly supplying a cumulative spatial representation, complementing expressibility analyses and studies of learned computation.

Spatial text representations. Chargrid and layout-aware document models use spatial grids or boxes (Katti et al., 2018; Xu et al., 2020a; 2021); rendered-text encoders learn from text images (Rust et al., 2023; Lotz et al., 2023). Text-art synthesis and perception (Xu et al., 2015; Akiyama, 2017; Jia et al., 2026) and sub-token understanding (Hou et al., 2026) offer complementary views of spatial glyph structure. Training on ASCII layout construction also improves spatial reasoning in language models (Huang et al., 2026). Here a completed source row and fixed local advances determine source positions without sample-specific bounding boxes or image features; the controlled intervention concerns this computable representation in conditional prediction.

Scope of the evidence. The evidence covers MS PGothic text art and observed-history nonspace prediction, with three paired fits per stage and an additional three-pair stage for normalization. The main ordinary additions were reused across follow-up studies; the screened event evaluation uses newly acquired works from one event, and stress targets form a separately selected, enriched population (Appendix L). The larger normalized check changes capacity and regularization jointly; raw/augmented causal models are separate tests. The gains may reflect both accumulation and source–target metric consistency; comparisons with learned cumulative mechanisms such as CoPE, broader training searches and other rendering settings remain open.

7 Conclusion

Explicit cumulative position improves conditional prediction when its local measurements are already supplied and canonicalization guarantees invariance to the specified whitespace rewrites. Fixed-recipe repeats, representation controls and a larger regularized normalized model support this finding; frozen models retain the benefit on screened event artwork. For the tested proportional-text predictors, computability and encoding invariance are therefore insufficient reasons to treat an explicit relational feature as redundant.

REPRODUCIBILITY STATEMENT

The original-stage entry supplies six models and a builder matching all 18 array hashes; the four-arm entry supplies twelve models, frozen features and replay. The normalized pseudo-metric and residual-dropout entry supplies twelve further models with separate statistical replay and an original-Unicode input bridge. The larger normalized entry adds six selected models and block-statistic replay, reusing the same input builder. The included record manifest and extractor recover all 2,951 HukuTemp records from the specified public archive and rebuild all 18 numerical arrays with their original hashes. Separately, `kouhaku-reproduce/` reconstructs all 19 event input arrays from public pages and supplies scoring and replay. Across these entries, checks on 512 fixed natural targets per model reproduce all archived character choices with NLL discrepancies below 2×10^{-5} bits. The event entry additionally passes a clean public-source full-forward check: all 148,488 character choices match, with NLL differences below 5×10^{-5} bits. Historical population selection and model training are not rerun (Appendix L.5).

DATA AND ETHICS STATEMENT

The community text art comes from HukuTemp and public AA competition pages. HukuTemp's README permits redistribution and reposting of "this information" (Appendix H.1); this does not verify every artwork's or third-party character's rights. The package excludes the full natural corpus, new event artwork text and proprietary font binary. The forward example uses self-authored inputs and fixed advances. The task predicts characters without inferring creator identities or personal attributes; source-file and visible-signature grouping handles detected dependence without establishing independent authorship.

AI USE STATEMENT

AI tools assisted with literature retrieval and discovery, research and experimental-design feedback, code implementation and debugging, input preprocessing and reconstruction, statistical checks and interpretation of results, drafting and revising manuscript sections, and figure-generation code. The authors are responsible for the paper's content, analyses and conclusions.

REFERENCES

1anon8anon1. Japanese AA (Shift_JIS art): MLT collection v32.0. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/1anon8anon1/yaruo-mlt-aa. Revision a95fa076383f788b38dc76ecd5f6900313d6e3cd; archived file identities verified September 8, 2026.

Osamu Akiyama. ASCII art synthesis with convolutional networks. In *NIPS 2017 Workshop on Machine Learning for Creativity and Design*, 2017. URL https://nips2017creativity.github.io/doc/ASCII_Art_Synthesis.pdf.

Anders Carlsson. Beyond encoding: A critical look at the terminology of text graphics. *WiderScreen*, 20(1–2), 2017. URL https://widerscreen.fi/assets/Anders-Carlsson-1-2-2017.pdf.

Hanseul Cho, Jaeyoung Cha, Pranjal Awasthi, Srinadh Bhojanapalli, Anupam Gupta, and Chulhee Yun. Position coupling: Improving length generalization of arithmetic transformers using task structure. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.20671.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important, 2024. URL https://arxiv.org/abs/2405.18719.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

Shuyang Hou, Yi Hu, and Muhan Zhang. SubTokenTest: A practical benchmark for real-world subtoken understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19957–19999. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.915. URL https://aclanthology.org/2026.acl-long.915/.

Shiyuan Huang, Li Liu, Jincheng He, and Leilani H. Gilpin. Learning to draw ASCII improves spatial reasoning in language models, 2026. URL https://arxiv.org/abs/2604.14641.

Qi Jia, Xiang Yue, Shanshan Huang, Ziheng Qin, Yizhu Liu, Bill Yuchen Lin, Yang You, and Guangtao Zhai. ASCIIEval: Benchmarking models' visual perception in text strings via ASCII art. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/63f5c95b1e6364c42075f913d84ccb73-Paper-Conference.pdf.

Anoop R. Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4459–4469. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1476. URL https://aclanthology.org/D18-1476/.

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. doi: 10.52202/075280-1082. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4e85362c02172c0c6567ce593122d31c-Paper-Conference.pdf.

Kōhaku AA Battle contributors. The 22nd Kōhaku AA Battle: Original submission thread, 2026. URL https://bbs.yaruyomi.com/test/read.cgi/ban/1765010078/. Artwork submissions and author corrections, December 2025–January 2026. Accessed September 25, 2026.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1807.03247.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. Text rendering strategies for pixel language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10155–10172. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.628. URL https://aclanthology.org/2023.emnlp-main.628/.

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and Tom Goldstein. Transformers can do arithmetic with the right embeddings. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.17399.

Mona Font Project. Mona Font, 2005. URL https://monafont.sourceforge.net/index-e.html. Project documentation; last modified October 2005. Accessed September 13, 2026.

Monagiko AA Archive. The 22nd Kōhaku AA Battle: Red and white artwork catalogs, 2026. URL https://w.atwiki.jp/monagiko2/pages/1697.html. Red catalog, with the white catalog at https://w.atwiki.jp/monagiko2/pages/1698.html. Accessed September 25, 2026.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FkSp8VW8RjH.

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://www.sciencedirect.com/science/article/pii/S0925231223011864.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090. PMLR, 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

Xuemiao Xu, Linyuan Zhong, Minshan Xie, Jing Qin, Yilan Chen, Qiang Jin, Tien-Tsin Wong, and Guoqiang Han. Texture-aware ASCII art synthesis with proportional fonts. In *Proceedings of the Workshop on Non-Photorealistic Animation and Rendering (NPAR '15, part of Expressive 2015)*, pp. 183–193. The Eurographics Association, 2015. doi: 10.2312/exp.20151191. URL https://diglib.eg.org/items/a1f20aa5-8855-4d76-ba34-cefd9ed61be9.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multimodal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2579–2591. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.201. URL https://aclanthology.org/2021.acl-long.201/.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1192–1200. ACM, 2020a. doi: 10.1145/3394486.3403172.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=r1eBeyHFDH.

Yaruo-thread MLT Collection. HukuTemp: Community mlt archive, releases 32.0–32.2, 2026. URL https://bbs.yaruyomi.com/test/read.cgi/ban/1759016679/. Release dates: June 27, July 26, and September 6, 2026. Collection thread and public distribution links. Accessed September 24, 2026.

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.

SUPPLEMENTARY GUIDE

Table 7 compares the designs supporting the current W claim. New initializations, newly trained models and newly acquired material are distinct evidence changes. All ordinary comparisons reuse the previously examined HukuTemp additions. The initial frozen event evaluation adds newly acquired, screened artworks. The later event controls reuse that event frame after its Index/Scan result was known.

Table 7: Current W evidence at a glance. "All" means canonicalization of every run in the specified whitespace equivalence family; "raw/aug." means raw inputs with one-step recoding augmentation. Bi/C denote bidirectional/causal source encoding; target history remains causal. A/B are the AdamW recipes defined below; p is residual dropout. This is a design map, not a pooled comparison or an additional inference family.

Comparison and finding	Input / encoder / recipe	Scored endpoint	Fits and seed reuse	Evaluation material
Original normalized W: gain after exact invariance (A)	All; Bi $64/2$; A, $p = 0$	Common Dev choice: epoch 8	Three new pairs; earlier S0 identities	Ordinary, reused
Fixed-endpoint repeat: gain across new initializations (A.1)	All; Bi $64/2$; A, $p = 0$	Fixed epoch 8	Three new pairs and new S1 identities	Same ordinary
Width-only / Linear: widths alone and tested row scale do not match Scan (A.2)	All; Bi $64/2$; A, $p = 0$	Fixed epoch 8	Six new controls using S1 states/orders; existing Index/Scan	Same ordinary
Pseudo / Scan: true geometry beats the specified pseudo-metric (B.1)	All; Bi $64/2$; A, $p = 0$	Fixed epoch 8	Both arms retrained; S1 states/orders reused	Same ordinary
Dropout Index/Scan: gain with residual regularization (B.2)	All; Bi $64/2$; A, $p = .1$	Fixed epoch 8	Both arms retrained; S1 states/orders reused	Same ordinary
Larger normalized W: gain under joint capacity/recipe change (B.3)	All; Bi $128/4$; B, $p = .1$	Common Dev choice: epoch 6	Three new pairs and new S2 identities	Same ordinary
Frozen event Index/Scan: gain on screened new works (A.3)	All; Bi $64/2$; A, $p = 0$	Existing epoch-8 models	Fitted S1 Index/Scan reused; no training	New event; rules fixed before scoring
Event Width-only / Linear: controls also fall short of Scan (A.4)	All; Bi $64/2$; A, $p = 0$	Existing epoch-8 models	Fitted S1 controls/Scan reused; no training	Same event, already examined
Larger causal W: gain in selected larger systems (C)	Raw/aug.; C $128/4$, $256/4$; $p = 0$	128: Index B/4, Scan A/8; 256: both B/4	New fits at each size; S0 identities reused	Ordinary; stress secondary, both reused
W recoding / mask studies: lower mean recoding sensitivity (D, E, C)	Raw/aug.; C/Bi $64/2$, Bi $128/4$; A, $p = 0$	Common Dev epoch 8 in each setting	S0; new fits per setting; mask follow-up reuses initial states	Ordinary; stress frozen initially, then reused

Reading the design map. Recipe A uses learning rate 0.001 and weight decay 0.0001; B uses 0.0003 and 0.001. The notation d/ℓ gives width and layer count. S0 denotes seeds 20260919/20/21, S1 denotes 2026092401/02/03 and S2 denotes 2026092501/02/03. A reused seed identity does not imply identical weights across architectures. Ordinary material has 2,951 records and 611 blocks; event material has 19 works and 18 visible-signature groups; stress material has 512 components. These populations and their weights remain separate. The exact selection rules, per-pair results and intervals follow in the linked sections. Table 47 separately preserves the original primary endpoints and interval families, including the original K_2 designation of the recoding studies.

Where to read next. For the central normalized comparison, controls, event evaluation and descriptive spread, start with Appendix A. Appendices B–C cover the additional training settings; Appendices D and E give the unnormalized behavioral studies. Earlier F-series comparisons and their controls are grouped in Appendix F and Appendices I–K. The shared model and material definitions are in Appendices G–H. Appendix L records inference scope and historical boundary results; Appendix L.5 maps the current executable package to these results.

Current package and historical locators. The top-level `README.md` is the command entry point. Appendix L.5 identifies the bundled executable entries and their external dependencies. Other artifact paths in the earlier W/F, model/material, training, recoding and descriptive sections refer to the separately archived legacy packet; its scripts, ZIP files and companion PDFs are provenance locators, not current bundled files or download addresses. The current W descriptive entry is explicitly identified in Appendix A.5.

A ALL-RUN NORMALIZATION WITH SHARED LOCAL WIDTHS

Intervention and exact scope. This separately trained follow-up combines shared true local widths, bidirectional source encoding, and normalization of every maximal source run containing only U+0020/U+3000, including leading, internal, trailing, and all-space rows. In the fixed MS PGothic metric, these spaces advance 5 and 11 pixels. For run advance W, the normalizer chooses nonnegative integers (a, b) minimizing $(a + b, a, b)$ lexicographically subject to $5a + 11b = W$, then emits b U+3000 characters followed by a U+0020 characters. This preserves total advance and nonspace-character pen origins. It does not normalize every Unicode whitespace character or every visual equivalence. Only source rows change: target history, legal query, labels, target IDs, record denominators, and original weights remain fixed.

Both arms use the two-layer, four-head, $d = 64$ W architecture with 348,811 parameters, shared source/history encoder weights, causal target history, complete source access, zero dropout, and an active bias-free $2 \to 64$ source projection. They receive the same normalized local width, including true widths for source characters whose IDs map to UNK. The second scalar is a moment-matched source index or cumulative true advance. The target query scale remains the original fit-only scale. The fit, development and ordinary-additions source banks all undergo the same normalization.

Table 8: All-run source preprocessing. Source-bank rows are storage units, not independent authors. Every original target remains included.

Pool	Targets	Source rows	Changed rows	Positions: raw → normalized
Fit	537,584	16,521	10,294	958,590 → 958,530
Development	549,663	17,165	11,087	1,002,270 → 1,002,192
Additions	2,662,587	83,479	54,155	4,938,406 → 4,938,100

Many rows change space order, but the position-count reduction is small; this is not a large sequence-compression intervention. Empty-source rows (57/35/149 in fit/development/additions) and output-UNK targets remain. All five raw/left/right rewrite views in Appendix D map to one source input. Their deterministic prediction identity is a preprocessing property; this experiment has no learned-JS robustness endpoint.

Fit-only calibration. Let w_j denote the positive, uncentered true Unicode advance and j the zero-based index in the normalized source. The features are

$$\widetilde{w}_j = (w_j - m_w)/s_w, \quad g_j = \frac{\sum_{t<j} w_t + 0.5 - \mu_q}{\sigma_q}, \quad h_j = m_g + \frac{s_g}{s_j}(j - m_j).$$

Here $(\mu_q, \sigma_q) = (281.0954529896723, 164.0111223765064)$; $(m_w, s_w) = (6.637306723641635, 4.095337485834411)$; $(m_g, s_g) = (-0.020200379993929, 1.124950864334976)$; and $(m_j, s_j) = (42.710315986576596, 32.058772408151626)$. Moments use valid source-position occurrences weighted by the multiplicity of fit targets accessing each row. The four scheduled exposures

coincide after normalization and give 157,540,996 valid position occurrences per cycle. PAD and all development/evaluation positions are excluded. Population standard deviations and affine formulas are computed in float64, then features are cast to float32; PAD features are zero. Width accumulation precedes normalization.

Original training and selection. Seeds 20260919/20260920/20260921 pair identical untrained weights and target orders across arms; no fitted weights are reused. Each trajectory completes 20 epochs, 168,000 updates and 10,751,680 target exposures, with batch size 64 (tail 48), FP32, AdamW learning rate 10^{-3} and weight decay 10^{-4}. The former raw/raw/left1/right1 cycle becomes four identical normalized views without extra exposures. A new common endpoint minimizes the equal mean of six weighted normalized-development NLLs over epochs 4/8/12/16/20, with exact ties going earlier. Epoch 8 (67,200 updates) is selected; it is not inherited from a previous experiment. Table 9 reports every candidate and fit, including later deterioration.

Table 9: All five all-run development candidates (bits/target). Seed suffixes 19/20/21 denote 20260919/20260920/20260921. The six-model mean selects epoch 8.

Epoch	Index 19	Scan 19	Index 20	Scan 20	Index 21	Scan 21	Mean
4	3.483408	3.456121	3.484079	3.461766	3.479721	3.456825	3.470320
8	3.463593	3.437515	3.489019	3.448868	3.461905	3.417078	3.452996
12	3.485615	3.431797	3.507578	3.452925	3.482385	3.435569	3.465978
16	3.543975	3.473407	3.551811	3.515878	3.518791	3.472977	3.512807
20	3.573264	3.518556	3.585739	3.555300	3.549930	3.521337	3.550688

Original ordinary prediction and uncertainty. The original stage's single prespecified primary endpoint is the equal-seed mean of original-weighted NLL(index)−NLL(scan), in bits/target. The previously examined additions retain all 2,951 records and 611 original source-file blocks. Each target receives its record's original weight divided by that record's original target count, including UNK and empty-source targets. Table 10 reports all original fits and both absolute metrics.

Table 10: Original all-run ordinary results at the common epoch 8. Accuracy is descriptive.

Seed	Index NLL	Scan NLL	NLL gain	Index acc. (%)	Scan acc. (%)
20260919	3.436184	3.417732	0.018452	51.4247	51.6415
20260920	3.461365	3.415465	0.045900	51.1565	51.6403
20260921	3.437513	3.392260	0.045252	51.4342	51.7780
Mean	3.445020	3.408486	0.036535	51.3384	51.6866

The primary gain is 0.036534584 bits/target with the original material 95% percentile interval [0.033186389, 0.040122229]. The 10,000 paired PCG64 draws (seed 20260923) resample all 611 blocks with replacement after averaging the three fixed fitted pairs, preserving their common block multiplicities and recomputing the weight denominator in every draw; NumPy linear quantiles supply the endpoints. Accuracy improves by 0.348144 percentage points and remains descriptive. The seed-difference sample SD is 0.015664 bits. A later descriptive Student-t interval across three paired differences is $[-0.002376, 0.075445]$ (95%, $df = 2$). Its normal-difference assumption cannot be checked with three fits. The positive material interval therefore does not establish that a newly trained pair will have a positive gain, and the seed interval is not an adjustment for the shared endpoint selection or adaptive research process.

A.1 Additional fixed-recipe training repeats

After the original result, a separate protocol fixes three new paired seeds (2026092401/2026092402/2026092403), the existing training recipe and epoch 8 before any additional fit. Every new model stops at 67,200 updates (4,300,672 target exposures); its development score is diagnostic only. Architecture, optimizer, canonical inputs, feature calibration, target identities and weights remain unchanged. Within each new pair, both arms start from identical

untrained tensors and use the same eight target permutations. All six planned fits and complete ordinary evaluations finish without replacement or retry.

The additional stage's primary endpoint is the NLL difference averaged over these three new pairs; accuracy is secondary. It uses all 2,662,587 targets and the same previously examined 611 blocks. The analysis retains every new seed and uses the original block-bootstrap procedure. Original fits are excluded from this stage's primary estimate because they participated in selecting epoch 8. No six-pair pooled primary or new-material confirmation is claimed.

Table 11: Additional normalized W fits at fixed epoch 8. Positive gains favor W-scan. Each seed pairs the initial tensors and target orders.

Seed	Index NLL	Scan NLL	Gain	Index acc. (%)	Scan acc. (%)
2026092401	3.442152	3.393232	0.048919	51.226898	51.715668
2026092402	3.434024	3.387429	0.046595	51.535830	51.963316
2026092403	3.447528	3.419913	0.027615	51.117275	51.437869
Mean	3.441235	3.400192	**0.041043**	51.293334	51.705617

The gain is 0.041043097 bits/target, with a conditional material 95% interval [0.037927284, 0.044240997]. Accuracy improves by 0.412283 percentage points. The sample SD of the three paired NLL differences is 0.011687 bits; their descriptive mean t interval is [0.012010030, 0.070076164] ($df = 2$). This interval describes the mean under the stated fit-difference assumptions, not the outcome of a future pair. Independent reaggregation of all six per-target score vectors reproduces the reported endpoints and all 10,000 bootstrap replicates. Checkpoint receipts, optimizer steps, paired initial states and all eight data-order hashes pass separate integrity checks. The anonymous summaries, protocol and aggregate replay inputs are in `supplementary/new-results/allrun-repeat/`; this package omits the new checkpoint bodies and natural per-target score vectors. The separate `representation-controls/` entry now supplies the selected Index/Scan parameter dictionaries alongside the new control models (Appendix A.2).

Evidence boundaries and verification. The training stages above combine the three stated controls in one small bidirectional system and reuse ordinary material examined in earlier studies. The larger normalized comparison is reported separately in Appendix B.3; second-font, free-generation and internal-algorithm questions remain outside these experiments. The separate frozen-model event evaluation below addresses newly acquired material. Comparisons between the normalized models and separately fitted raw W models do not identify the causal effect of normalization: calibration, training views and endpoint selection were recomputed. Saved-prediction replay and independent target-to-block reaggregation exactly reproduce the original summary and bootstrap arrays. A separate input audit checked all 117,165 source rows and 6,898,822 normalized source positions, including true Unicode widths and exclusive prefix sums. Those checks execute no new training, model forward pass, or pixel rendering. The numerical records are in `supplementary/new-results/allrun/`; runnable core reproduction and external-asset boundaries are described in Appendix L.5.

A.2 Fixed-recipe representation controls

Questions and feature definitions. This follow-up distinguishes an additional index scalar, a coarse row-scale coordinate and the exact cumulative source coordinate. All arms retain the source characters, true Unicode widths looked up before UNK mapping, source-index sinusoidal encodings, target history and query, complete source access, and the bidirectional $d64/2$ architecture. Width-only supplies $[\widetilde{w}_j, 0]$ to the original bias-free $2 \to 64$ projection. The second column remains in the parameter dictionary but receives no data gradient: this is a natural feature-removal baseline with the same nominal parameter count, not a second active matched channel.

For a nonempty canonical source of length L, Linear constructs

$$\ell_j = 0.5 + \frac{j}{L} \sum_{t=0}^{L-1} w_t, \qquad \widetilde{\ell}_j = m_g + s_g \frac{\ell_j - m_\ell}{s_\ell},$$

and supplies $[\widetilde{w}_j, \widetilde{\ell}_j]$. The first expression spreads the known total source advance uniformly across indices; the second matches the existing normalized Scan channel's training mean and standard deviation. It retains row-scale information under a global affine calibration, rather than providing exact prefix positions or raw pixel coordinates. The full source row, including its final glyph, is already available. This feature does not use a target character or unobserved target suffix. Empty sources use the shared NULL path.

The calibration uses only canonical fit-source positions, weighted by the number of fit targets referencing each row. PAD and empty sources are excluded. Population moments are computed in float64 before casting features to float32. The frozen calibration has $(m_\ell, s_\ell) = (266.1031019633772, 186.43133794355958)$ and uses the existing (m_g, s_g). Source normalization, local-width calibration and target-query scaling remain unchanged.

Pairing and fixed training budget. The new controls use seeds 2026092401/2026092402/2026092403. Each starts from the byte-identical untrained state and RNG state used for the corresponding additional Index/Scan pair, with the same eight target-order permutations. No fitted weights initialize a new control. Both new arms retain FP32, batch size 64 (tail 48), zero dropout, AdamW learning rate 10^{-3} and weight decay 10^{-4}. Epoch 8 is fixed at 67,200 updates and 4,300,672 target exposures; development scores are diagnostic and do not reselect an endpoint. Existing fitted Index/Scan models and their ordinary scores are reused as separately hash-bound references.

Prespecified contrasts and uncertainty. The follow-up's single primary endpoint is the equally weighted mean NLL(Width-only)−NLL(Scan) across the three matched seeds. The prespecified secondary NLL(Linear)−NLL(Scan) contrast is descriptive; its interval does not supply joint confirmatory coverage with the primary contrast. Index−Width-only differences and accuracy for all arms are descriptive. All comparisons retain the original 2,662,587 ordinary targets, 2,951 records, 611 source-file blocks and hierarchical weights, including output UNK and empty sources.

Material 95% intervals use 10,000 shared paired block draws, PCG64 seed 20260923, after averaging the three fitted-seed differences. Every draw recomputes the original weight denominator; percentile endpoints use NumPy linear quantiles. Per-seed differences and descriptive mean Student-t intervals ($df = 2$) remain separate. The latter rely on a normal-difference assumption and are not prediction intervals for a new fit. No additional significance, stress or subgroup tests belong to this follow-up.

Evidence identity and completion rule. The control design was fixed after earlier findings but before these new fits. It reuses the additional stage's fitted Scan references, paired seed identities and examined ordinary material; it is a diagnostic follow-up, not new-material confirmation or an additional independent repeat of Scan. The original selected-endpoint stage, the subsequent fixed-endpoint repeat stage and this control study retain separate estimates and protocol identities. The original stress studies' primary K_2 endpoints and intervals are unchanged; the controls add no K_2 comparison.

Inclusion requires all three complete scores for the relevant new arm. The predeclared time/compute stopping rule permits reporting the complete Width-only primary if Linear is incomplete. Any partial Linear outputs would remain archived, without being presented as a complete three-seed comparison. Neither endpoints nor seed, arm or population inclusion are chosen from the new predictive outcomes.

Results and completed scope. All six new models complete the fixed endpoint and ordinary evaluation. Table 12 retains every paired result. Scan improves on Width-only by 0.060137791 bits/target, with material 95% interval [0.056611179, 0.063721830] and descriptive mean fit interval [0.051413482, 0.068862100]. The secondary Linear contrast is 0.036775380 bits/target, with material 95% interval [0.034423369, 0.039140666] and descriptive mean fit interval [0.005719604, 0.067831156]. These are separate marginal summaries, without joint coverage. Scan's accuracy improvements are 0.647024 and 0.314390 percentage points, respectively. The Index−Width-only NLL differences are -0.013359, -0.015455 and -0.028470 bits; removing the extra index scalar does not improve this baseline under the fixed recipe.

Table 12: Fixed-recipe representation controls on the canonical ordinary population. Index and Scan are the existing fixed-endpoint references; Width-only and Linear are newly trained from the same initial states. Positive differences favor Scan.

Seed	Width NLL	Index NLL	Linear NLL	Scan NLL	Width−Scan	Linear−Scan
2026092401	3.455511	3.442152	3.427835	3.393232	0.062278	0.034602
2026092402	3.449479	3.434024	3.437650	3.387429	0.062050	0.050221
2026092403	3.475998	3.447528	3.445416	3.419913	0.056085	0.025503
Mean	3.460329	3.441235	3.436967	3.400192	0.060138	0.036775

Seed	Width accuracy (%)	Index accuracy (%)	Linear accuracy (%)	Scan accuracy (%)
2026092401	50.933794	51.226898	51.343146	51.715668
2026092402	51.280934	51.535830	51.453937	51.963316
2026092403	50.961053	51.117275	51.376600	51.437869
Mean	51.058594	51.293334	51.391228	51.705617

Independent block-level reaggregation reproduces the earlier Index/Scan result and both new contrasts, including all 10,000 bootstrap replicates. Model and score bindings, initial tensors, RNG state, target permutations and the frozen feature calibration pass integrity checks. The `representation-controls/` entry supplies the twelve selected parameter dictionaries, calibrated feature implementation and block-level statistical replay. Natural source inputs retain the external-data boundary described in Appendix L.5.

A.3 Frozen-model evaluation on screened event material

This evaluation asks whether the additional fixed-endpoint models retain an Index–Scan prediction difference on newly acquired, screened artwork. The material, six models, preprocessing and analysis were fixed before model inference. It uses existing fits and a separate population, rather than adding training repeats or enlarging the ordinary-additions estimate.

Sources, extraction and qualification. We froze all 49 single-picture entries in the 22nd Kōhaku AA Battle's red/white catalogs and checked their text against the original submission thread (Monagiko AA Archive, 2026; Kōhaku AA Battle contributors, 2026). The event spans December 2025–January 2026. Titles and commentary are excluded by recorded artwork boundaries; multi-post continuations and explicit author corrections are resolved before scoring. Raw snapshots preserve original-post whitespace for the audit. Model text follows the inherited preprocessing: trim each line's trailing whitespace and remove outer empty lines, preserving internal spacing and line breaks. Source URLs, post identities, boundaries, correction choices and text hashes define the extraction contract; material was not selected by awards, predicted performance or output-vocabulary coverage.

Conservative prior-material screening quarantines 29 entries with detected relations to archived material. It compares exact normalized texts and whitespace-free texts, normalized long-line sets at Jaccard ≥ 0.5, and recorded containment relations against the three public HukuTemp versions and bound usage metadata. Archive presence is not itself a claim of training leakage. One additional entry, white-24, contains U+58DE, absent from the fixed advance table, and is excluded without extending that table. The remaining 19 entries all satisfy the inherited shape range (at least 15 rows, at least 16 characters in the widest row, and no TAB); those shape thresholds are material-scope rules, not model limits. This finite screen reduces detected overlap; it does not certify the absence of every crop, modified glyph or uncredited template. Reserved material remains sealed; only its existing identity metadata is used for exclusion.

Catalog artwork containers declare MS PGothic first at 16px. The original thread uses a Saitamaar-first override with MS PGothic among its fallbacks. These declarations and support in the frozen advance table establish input compatibility within the same AA tradition, not each author's actual font or exact browser advances. No new font, rendering metric or pixel-identity claim is introduced for this evaluation.

Inputs, models and weights. The 19 works contain 24,748 targets across 728 evaluated rows. All 310 output-OOV targets retain the original probability-floor treatment. Source characters mapped to UNK retain their real Unicode advances; the source bank has 455 such occurrences. Canonicalization changes 571 rows while preserving nonspace pen origins; raw and canonical banks both contain 52,608 character positions. The longest source has 203 characters. One empty-source row, accounting for 12 targets, follows the shared NULL rule; PAD positions contribute neither width nor calibration statistics. Target history and query are built from the original observed prefix. Independent input checks cover all target/prefix boundaries and verify that changing the target or future suffix leaves the prediction input unchanged.

The initial frozen comparison evaluates the additional-stage Index/Scan models with seeds 2026092401/02/03: bidirectional $d64/2$, epoch 8, step 67,200. Weights, vocabulary, widths, calibration and all preprocessing remain fixed; there is no new training, endpoint selection or calibration fit. All six complete FP32 forward passes use the same numerical backend and batching rules. Their complete score files and before/after model hashes are sealed and verified before aggregate analysis.

Visible submission signatures define 18 groups. White-18 and white-23 share one signature and form a single group; all other works form singleton groups. These are observable grouping labels, not verified natural-person identities or independent-author certificates. For group g with m_g works, the frozen estimator is

$$\widehat{L}_a^{\text{event}} = \frac{1}{3 \cdot 18} \sum_{s=1}^{3} \sum_{g=1}^{18} \frac{1}{m_g} \sum_{i \in g} \frac{1}{T_i} \sum_{t=1}^{T_i} -\log_2 p_{a,s}(x_{i,t}). \tag{5}$$

Thus targets are equal within each work, works equal within each group, and groups and seeds equal. The sole primary contrast is Index minus Scan NLL; accuracy uses the same weights and is descriptive. A paired group bootstrap draws 18 groups with replacement in each of 10,000 replicates (PCG64 seed 20260923), using the same draws for both arms and every seed. Its percentile 95% interval conditions on these fitted models and this screened, nonprobability event sample. A separate descriptive Student-t interval ($df = 2$) summarizes the mean of the three paired differences. It is not a prediction interval for a new fit; neither interval establishes cross-event or independent-author coverage. No new significance test, subgroup search or pooled ordinary/event estimate is introduced.

Results and execution checks. Table 4 reports the primary gain of 0.0764 bits/target and conditional material interval [0.0457, 0.1099]. All three existing fitted pairs favor Scan on this material, while their descriptive mean interval [−0.0445, 0.1972] crosses zero. Table 13 gives every pair and both absolute metrics; Table 14 retains every artwork, including negative individual differences. The weighted accuracy changes from 36.097% to 36.984%, a descriptive increase of 0.886 percentage points. The source population and its weighting differ from the ordinary additions, so their absolute losses and gap magnitudes are not a controlled between-population comparison.

Table 13: All existing model pairs on the screened event population. NLL is in bits/target, accuracy in percent, and positive differences favor Scan. The mean uses the fixed group–work–target weights and averages the three fits; it does not treat them as additional material groups.

Seed	Index NLL	Scan NLL	NLL gain	Index acc.	Scan acc.	Acc. gain (pp)
2026092401	4.8825	4.7983	0.0842	35.950	37.031	1.080
2026092402	4.8646	4.7440	0.1206	35.949	37.038	1.090
2026092403	4.8619	4.8376	0.0243	36.392	36.882	0.490
Mean	4.8697	4.7933	0.0764	36.097	36.984	0.886

Reconstruction and evidence scope. The `kouhaku-reproduce/` entry supplies the frozen public-source locators, extraction contract, model references, input builder, forward scorer and statistical replay. Reconstruction from either the recorded forum posts or catalog pages matches all 19 artwork hashes; the builder matches all 19 input arrays and the prepared NPZ identity. A clean-directory online download and rebuild independently reproduces these inputs. On 512 fixed targets per model, CPU checks reproduce all 3,072 original CUDA character choices with maximum NLL discrepancy 1.9606×10^{-5} bits; these retain their original subset scope. A subsequent clean-directory check retrieves the public forum pages again, reconstructs every input and evaluates all

Table 14: Complete artwork-level results, without inferential subgroup claims. R/W identify the red/white catalog and preserve its entry numbers. NLL and accuracy columns average the three existing fits within each work; the last three columns give every paired NLL difference. W18 and W23 share one group and each receives half that group's primary weight. Full-precision per-fit absolute scores and accuracies accompany the replay data.

Work	Targets	Index NLL	Scan NLL	Acc. I/S (%)	Δ_{01}	Δ_{02}	Δ_{03}
R01	565	5.0438	4.9210	31.09/31.74	0.1002	0.1475	0.1206
R04	1,893	4.0636	3.9887	34.58/36.10	0.1715	0.0437	0.0094
R09	226	6.9068	6.7250	28.76/31.56	0.2747	0.1160	0.1548
R10	2,879	6.5656	6.5753	22.16/21.98	−0.1163	0.0982	−0.0110
R11	453	5.8861	5.8447	25.09/26.34	0.1041	−0.0279	0.0482
R15	519	4.5334	4.5133	30.19/30.89	0.0425	0.0159	0.0018
R17	425	3.5823	3.5202	53.80/54.35	−0.0030	0.0702	0.1193
R20	562	4.6159	4.5638	25.68/26.51	0.0751	0.0353	0.0458
R22	1,145	3.4086	3.4012	49.32/50.28	−0.0317	−0.0154	0.0692
W02	2,744	1.6681	1.5299	72.52/73.64	0.0695	0.1718	0.1731
W05	2,127	5.1788	5.0828	45.09/45.81	0.1272	0.0701	0.0905
W08	1,463	4.4132	4.3768	40.90/41.22	0.0664	0.0444	−0.0015
W09	2,198	6.5175	6.2853	24.42/26.77	0.1185	0.9252	−0.3469
W12	1,045	3.5156	3.4929	50.56/50.94	0.0513	0.0309	−0.0139
W16	296	5.8091	5.7876	32.66/33.56	0.0596	0.0374	−0.0327
W17	2,075	4.3513	4.3123	33.67/34.46	0.0074	−0.0107	0.1203
W18	1,225	6.8882	6.8794	14.37/14.29	0.0224	0.0238	−0.0200
W21	1,570	5.5870	5.3810	26.65/26.62	0.3641	0.3820	−0.1280
W23	1,338	5.1269	5.0766	30.87/31.59	0.0453	0.0489	0.0565

24,748 targets for each of the six frozen models on CPU. All 148,488 character choices match the original CUDA run, with maximum NLL difference 4.5920×10^{-5} bits, within the declared 10^{-4}-bit tolerance. Each model's weighted mean differs by less than 3.45×10^{-7} bits. This repeats software execution on the same frozen frame, not material selection. Independent reaggregation reproduces the primary result and bootstrap interval. Natural artwork text, HTML and prepared arrays are obtained through the documented source path rather than redistributed in the package. Source changes fail explicit identity checks. Appendix L.5 distinguishes this entry from the HukuTemp reconstruction requirements. The result extends the fixed-model prediction finding to screened material acquired after the preceding research decisions, within one AA event and rendering tradition. It provides neither new training repetitions nor cross-font or independent-author validation.

A.4 Event representation controls and influence checks

Fixed models, subsequent comparisons. After observing the initial event Index/Scan result, we scored the three existing Width-only and three existing Linear models on exactly the same 19 artworks. Each model retains its additional-stage seed, epoch-8 endpoint, weights and training-only calibration. No target, group, weight or input feature was selected or refitted. The protocol specified Width-only−Scan NLL as the primary follow-up contrast and Linear−Scan as a secondary descriptive contrast before these six forward runs. The original event Index/Scan result remains a separately frozen comparison; this extension is a diagnostic on examined material.

All six new full-population CPU score files were sealed and their stopped worker and file identities verified before aggregation. The original Index/Scan CUDA scores remain separately bound references. Independent reaggregation recovers their original estimates to numerical precision. All four arms use Equation 5's weights. Paired group intervals use 10,000 PCG64 draws with seed 20260923 and the original fixed lexicographic group order, retained explicitly in the replay inputs. Each nominal 95% material interval is marginal; the secondary interval does not establish joint coverage. Accuracy is descriptive.

The mean Width-only−Scan difference is 0.1034 bits/target, with conditional group interval [0.0461, 0.1786] and descriptive fit interval [0.0492, 0.1576]. Linear−Scan is 0.0592, with corresponding intervals [0.0296, 0.0933] and [−0.0166, 0.1349]. All three differences are positive for each contrast

Table 15: Every added control fit on the complete screened event frame. Positive NLL differences favor the same-seed frozen Scan model.

Control	Seed	NLL	Accuracy (%)	NLL − Scan	Acc. gain of Scan (pp)
Width-only	2026092401	4.9079	36.225	0.1095	0.805
Width-only	2026092402	4.8655	36.630	0.1215	0.409
Width-only	2026092403	4.9168	35.695	0.0792	1.187
Linear	2026092401	4.8710	36.076	0.0726	0.955
Linear	2026092402	4.8246	36.049	0.0806	0.989
Linear	2026092403	4.8619	36.452	0.0243	0.430

(Table 15). Thus the tested width-only and calibrated row-scale alternatives also leave an average cumulative-position benefit on this material. Full-precision scores for every work, fit and arm are retained in the extension's replay data.

Complete influence and weighting descriptions. A subsequent full-precision description of the original Index/Scan result finds positive mean gains for 18 of 19 works and 17 of 18 visible-signature groups. Removing each complete group in turn gives means between 0.0672 and 0.0814 bits/target; all three fit-specific overall gains remain positive after each removal. The full leave-one-group table is retained, including the group containing W09. This describes how the observed average is spread across the retained material, while the original per-work table continues to show individual fit reversals.

The original group-weighted mean is 0.0764 bits/target. Equal weighting of 19 works gives 0.0739, and equal weighting of all 24,748 targets gives 0.0799. These alternative averages and influence checks are post-hoc descriptions, introduce no new significance test, and do not replace the original population weights or interval. They use complete full-precision outputs rather than rounded table entries.

A.5 Distribution of the current all-run W prediction gain

This follow-up describes the selected all-run-normalized W-index/W-scan models in the original stage of Table 1: three paired fits at epoch 8, with bidirectional source encoding and shared true widths. It uses their saved ordinary per-target NLLs and accuracies, not new model scores. Its rules were fixed after the overall results and the earlier F analyses were known, but before inspecting the new W partitions. It introduces no new interval, test, model selection or independent confirmation. The complete outputs, fixed protocol, source hashes and replay entry are in the current packet's `supplementary/w-descriptive/`.

Full-population weights and block spread. Every one of the 2,662,587 targets, 2,951 records and 611 blocks retains its original primary weight. Reaggregation of all six per-target score vectors reproduces the archived block NLL and accuracy arrays bit for bit. For target weight ω_t and gain $d_{s,t}$, a block's contribution is $k_{s,b} = \sum_{t \in b} \omega_t d_{s,t}$ and its conditional gain is $k_{s,b}/m_b$, where $m_b = \sum_{t \in b} \omega_t$. Table 16 summarizes these block gains; numerical zero means absolute value at most 10^{-12} bits. There are 328 blocks with positive gains in every seed. Leaving out any single block gives an equal-seed overall gain between 0.035867 and 0.036777 bits. This is a material-influence diagnostic, not a confidence interval or a measure of training repeatability.

Table 16: Block-wise distribution of the current normalized W gain. The last row first averages the three gains within each block; it does not pool three runs into additional independent blocks.

Fit	Median gain	Positive	Negative	Numerical zero
20260919	0.017282	413	195	3
20260920	0.042853	513	95	3
20260921	0.040493	523	85	3
Mean within block	0.032656	537	71	3

Same-ordinal source–target position difference. For the canonical source of length L, define $\delta = |g_c - q|$ when $0 \le c < L$. Both coordinates are taken directly from the authenticated float64 pixel arrays, including their common $+0.5$ anchor. The source is canonicalized; the target column and query retain their original values. Thus δ compares positions at the same ordinal, not corresponding glyph identities, attention locations, or the two model-input scalars. The four fixed pixel strata are 0, $(0, 8]$, $(8, 16]$ and > 16; empty sources and nonempty sources with $c \ge L$ remain separate categories. Table 17 reports all six categories with their original masses and conditional absolute NLLs. Table 18 gives each fit and the contribution to the full-population gain. Category masses sum to one and contributions reconstruct each fit's original gain.

Table 17: **Normalized W gains span the input-defined offset strata.** Descriptive partition of Table 1, averaging its three fitted pairs. Mass is the original evaluation weight; NLLs and gains are conditional on each row. All targets remain, including undefined-offset categories.

Source–target offset	Mass (%)	Index NLL	Scan NLL	Gain
$\delta = 0$	7.34	3.18269	3.13010	0.05258
$0 < \delta \le 8$px	25.13	3.30130	3.27465	0.02665
$8 < \delta \le 16$px	18.16	3.40412	3.37713	0.02699
$\delta > 16$px	40.98	3.54979	3.50765	0.04214
Empty source	0.16	8.16128	8.20246	−0.04119
Nonempty source, $c \ge L$	8.22	3.59221	3.54511	0.04710

Table 18: Complete normalized W offset partition. Columns 2–4 are conditional gains by fit; the last column is the seed-averaged contribution with original population weights, in bits/target.

Offset	20260919	20260920	20260921	Contribution
0	0.066053	0.050754	0.040938	0.003861
$(0, 8]$px	0.005795	0.037798	0.036355	0.006697
$(8, 16]$px	0.009048	0.037456	0.034479	0.004902
> 16px	0.019960	0.052141	0.054317	0.017269
Empty	0.132340	−0.149869	−0.106035	−0.000068
Nonempty, $c \ge L$	0.025588	0.057784	0.057935	0.003874

Common-cell standardization. Source-length bins are 1–16, 17–32, 33–64, 65–128 and 129+; target-column bins are 0–7, 8–15, 16–31, 32–63 and 64+. A cell is retained only if each of its four defined-offset strata contains at least 20 targets, five records and three blocks. All 19 occupied cells meet this rule, retaining all 2,462,967 defined-offset targets and 91.6104% of the original population weight. Within each cell and stratum, original target weights give the conditional losses. Cell weights pool the original mass of all four strata and normalize over the retained cells; this same reference distribution is used for every stratum and seed. Undefined-offset categories retain their unadjusted results above.

Table 19: Normalized W results under common length/column cell weights. Absolute NLLs average three fits; all gains are descriptive, in bits/target.

Offset	Index NLL	Scan NLL	Gain: 20260919	20260920	20260921
0	2.895372	2.832804	0.090260	0.058291	0.039152
$(0, 8]$px	3.199031	3.169858	0.008307	0.039329	0.039882
$(8, 16]$px	3.357444	3.329335	0.009795	0.038659	0.035872
> 16px	3.653956	3.614698	0.015488	0.052559	0.049728

All four defined-offset strata have positive gains in each fitted pair, before and after this standardization. Their mean gains are not monotone in offset. In particular, coincidence of the two same-ordinal positions does not make the complete source's cumulative representation redundant. The partition describes one input relation; it neither holds all other source geometry fixed nor identifies an internal

calculation. The complete cell inventory, retention counts, absolute losses, accuracies and contrasts relative to the zero-offset stratum remain in the machine-readable report.

B WIDTH-KNOWN PSEUDO-COORDINATE AND RESIDUAL-DROPOUT DIAGNOSTICS

Two follow-up comparisons examine the supplied metric and the training recipe in the all-run-normalized W setting. The first compares Pseudo with Scan at zero dropout; the second compares Index with Scan at residual dropout $p = 0.1$. Each family trains both arms for three paired fits, giving twelve new fitted models. Every arm retains the true local Unicode width, complete source access, source-index sinusoidal encoding, target history and query, and the $d64/2$ bidirectional source encoder with 348,811 parameters. The first scalar is unchanged in both comparisons.

Shared fixed protocol. Seeds 2026092401/2026092402/2026092403 reuse the saved untrained tensors and CPU RNG states of the fixed-endpoint repeat stage. The eight training orders are reproduced from the frozen independent CPU stream and checked against their archived hashes. These are reused seed identities, with both arms retrained within each family; historical fitted Scan scores do not enter either contrast. FP32 AdamW uses learning rate 10^{-3}, weight decay 10^{-4} and batch size 64 (final batch 48). The endpoint is fixed at epoch 8, or 67,200 updates and 4,300,672 target exposures. The final development evaluation is diagnostic, without checkpoint reselection. The design is a follow-up to earlier findings, and evaluation retains all 2,662,587 previously examined ordinary targets, 2,951 records, 611 source-file blocks and original hierarchical weights, including output UNK and empty sources.

B.1 TRUE GEOMETRY VERSUS A FIXED PSEUDO-METRIC

Pseudo retains true local widths in its first channel but replaces the second channel with a cumulative coordinate from the frozen historical pseudo-width table. That table permutes positive nonspace advances within eight fit-frequency strata using a fixed SHA-based ordering (salt 20260917), preserves true whitespace advances, and uses a fixed hashed positive-width fallback for unseen nonspace characters. Both width lookups use original Unicode before mapping characters to UNK. For a canonical source row, the additional scalar is

$$p_j = 0.5 + \sum_{t<j} w^{\mathrm{pseudo}}(x_t), \qquad \widetilde{p}_j = m_g + s_g \frac{p_j - m_p}{s_p}.$$

The affine calibration matches the existing Scan channel's fit mean and standard deviation. Each valid source position is weighted by the number of fit targets referencing its row; PAD and empty sources do not enter the moments. The 537,584 fit targets give 39,385,249 valid-position occurrences, with $(m_p, s_p) = (428.675192, 329.880151)$ and $(m_g, s_g) = (-0.020200, 1.124951)$. Moments use float64 before one float32 feature cast; development data, ordinary evaluation data and target labels are not used in this calibration. Empty sources retain the shared NULL path.

The primary contrast for this diagnostic family is NLL(Pseudo)−NLL(Scan). Its mean is 0.041186 bits/target, with conditional material 95% interval [0.037586, 0.044519] and descriptive mean fit 95% interval [0.013559, 0.068813]. Scan improves accuracy by 0.4734 percentage points on average. Table 20 reports every pair. Thus the tested true rendered metric contributes beyond this positive, content-dependent monotone pseudo-coordinate when true local widths are shared and the specified whitespace encodings are normalized. The target query remains in the true rendered metric, so the contrast includes source–target metric consistency and does not identify an internal accumulation procedure.

B.2 INDEX VERSUS SCAN WITH RESIDUAL DROPOUT

The second family applies dropout $p = 0.1$ to self-attention outputs, feedforward outputs and the cross-attention retrieval output before their residual additions. Attention-probability dropout stays zero. Index and Scan share this recipe and its fixed endpoint. NLL(Index)−NLL(Scan) is 0.024199 bits/target, with conditional material 95% interval [0.021835, 0.026619] and descriptive mean fit

95% interval [0.013061, 0.035338]; Scan's accuracy advantage is 0.3268 percentage points. The representation gain therefore remains under this specified residual-dropout recipe. This within-recipe contrast does not estimate the causal effect of adding dropout to the earlier training runs.

Table 20: All paired fits in the two normalized diagnostic families. Each family has its own newly trained Scan arm. NLL is in bits/target; accuracy is in percent. Δ_{NLL} is control minus Scan and Δ_{acc} is Scan minus control in percentage points, so positive values favor Scan. Means give each of the three pairs equal weight.

Seed	Control NLL	Scan NLL	Δ_{NLL}	Control acc.	Scan acc.	Δ_{acc}
Pseudo versus Scan; residual dropout $p = 0$						
2026092401	3.435333	3.394736	0.040597	51.2610	51.7419	0.4809
2026092402	3.436668	3.384077	0.052590	51.4323	52.1169	0.6846
2026092403	3.451458	3.421087	0.030371	51.2195	51.4742	0.2547
Mean	3.441153	3.399967	0.041186	51.3043	51.7777	0.4734
Index versus Scan; residual dropout $p = 0.1$						
2026092401	3.412638	3.393616	0.019022	51.5941	51.7564	0.1623
2026092402	3.422964	3.396126	0.026838	51.6328	52.0258	0.3930
2026092403	3.435095	3.408357	0.026738	51.2350	51.6602	0.4252
Mean	3.423566	3.399366	0.024199	51.4873	51.8141	0.3268

Inference and reproduction. Each family has its own prespecified primary contrast and individual 95% interval; no joint coverage across the two families is claimed. Material intervals use 10,000 paired resamples of the 611 original source-file blocks with PCG64 seed 20260923. The three fixed pair differences are averaged before resampling; the same block multiplicities apply to all arms and seeds, and each draw recomputes its original weight denominator. Percentiles use linear quantiles. The separate descriptive mean Student-t intervals use $df = 2$ and a normal-difference assumption; they are not prediction intervals for a future fit. Accuracy remains descriptive. These comparisons reuse ordinary material and seed identities rather than adding independent material or seed confirmation, and they retain their own identities apart from the earlier representation controls and stress endpoints.

The accompanying `normalized-diagnostics/` entry separates statistical replay, model-forward execution and deterministic input reconstruction. The frozen Unicode records and reconstructed ordinary arrays supply the original-codepoint source sidecar through the input bridge; the bridge verifies original widths before UNK mapping and matches the full source-bank identity. Source acquisition and the original archive dependency remain as described in Appendix L.5. Independent full-score reaggregation reproduces both contrasts and all block statistics. On 512 fixed natural targets per model, the portable CPU implementation matches all 6,144 archived character choices, with maximum NLL discrepancy below 1.514×10^{-5} bits. This verifies a fixed forward subset; the input bridge separately checks all 4,938,100 source codepoints and all 2,662,587 targets against the original input identities.

B.3 Larger regularized predictors on fully normalized inputs

This follow-up crosses full source whitespace normalization with a larger bidirectional learner. Both arms retain true local widths, complete source access, target history/query, the same active two-scalar interface and the original all-run input calibration. The encoder has four blocks, dimension 128, four attention heads, feed-forward expansion four and residual-output dropout 0.1 (1,181,387 parameters). Target history remains causal.

The frozen recipe uses FP32 AdamW, learning rate 3×10^{-4}, weight decay 10^{-3}, batch size 64 (last batch 48), and eight complete epochs: 67,200 updates and 4,300,672 target exposures per model. Seeds 2026092501/2026092502/2026092503 share initial tensors and all eight data orders within each Index/Scan pair. Development candidates are epochs 1/2/3/4/6/8. Only after all six trainings finish does one common rule select the epoch with the lowest arithmetic mean of all six development NLLs, breaking exact ties toward the earlier epoch. Ordinary scoring follows that selection. The

common choice is epoch 6 (Table 22), not the earliest candidate. All candidates, final training states and scores are retained in the research archive.

Table 21: Complete larger normalized comparison at the shared epoch 6. Three paired fits on the same 2,662,587 ordinary targets and 611 blocks; NLLs are bits/target and accuracies are percentages.

Seed	Index NLL	Scan NLL	Difference	Index acc.	Scan acc.
2026092501	3.4078	3.3943	0.0135	51.507	51.879
2026092502	3.4174	3.3834	0.0341	51.496	52.000
2026092503	3.4083	3.3839	0.0243	51.650	51.934
Mean	3.4112	3.3872	0.0239	51.551	51.937

The sole primary contrast is Index−Scan ordinary NLL, retaining the original target–record–block weights and all outputs, including UNK. The conditional 95% material interval is [0.0200, 0.0273], from 10,000 paired block draws (PCG64 seed 20260923, linear quantiles). The three differences are 0.0135/0.0341/0.0243 bits; their sample standard deviation is 0.0103 and the descriptive mean t interval ($df = 2$) is $[-0.0017, 0.0495]$. Accuracy improves by 0.386 percentage points and is descriptive. No new subgroup, stress endpoint or pooled inference family is introduced.

Table 22: Common development selection for the larger normalized experiment. Every entry averages the same six candidate models. Selection uses no ordinary scores and applies one epoch to both arms and all three seeds.

Epoch	Six-model mean development NLL
1	3.5933
2	3.4982
3	3.4578
4	3.4374
6	3.4349
8	3.4438

This result supports predictive utility beyond the smallest normalized system. Capacity, dropout and optimization recipe change together, so cross-study differences do not isolate any one of them or establish a scaling law. The ordinary material was previously examined; new seed identities do not make it new material. The material interval conditions on the six selected fits and does not include retraining or selection uncertainty. The crossed setting is bidirectional, whereas the separate larger causal experiments use raw/augmented inputs.

The complete six-fit training, development and scoring pipeline took about 13.25 hours on one RTX 5060 Ti, including scoring 15,975,522 target–model combinations. After all workers exited, the archive was sealed and locally hash-verified. The frozen analyzer and an independently implemented full-score reaggregation agree on all six endpoints, common selection, paired differences and intervals. The `larger-normalized-reproduce/` entry supplies six unchanged FP32 selected-model dictionaries, frozen configuration and full block-statistic replay. Each exported model passes the fixed 512-target CPU check: all 3,072 character choices match archived CUDA scoring, with maximum NLL discrepancy below 8×10^{-6} bits. Portable batch tensors match the original scorer exactly. This is a subset implementation check, not a second full-population inference or a retraining reproduction.

C Development-selected causal capacity checks

Models, inputs, and scope. Two larger causal W configurations use $d = 128$, four layers and 1,181,387 parameters per arm, or $d = 256$, four layers and 3,924,555 parameters per arm. Both retain four attention heads, shared causal source/history encoding, full-source query access, 128-character history, head hidden size 128, zero dropout, and true Unicode local widths for both arms, including UNK source characters. W-index receives a moment-matched index; W-scan receives the

true exclusive cumulative advance. The same fixed-width table, target inputs, support and score floor apply as in Appendix D. The larger models retain raw source encodings and one-step internal-space augmentation. These causal models do not apply all-run normalization. The separate larger normalized bidirectional comparison is in Appendix B.3.

The fit/development arrays and their calibration are byte-identical to the original causal $d = 64$ inputs: 537,584 fit targets and 549,663 raw-development targets. The raw/raw/left1/right1 four-epoch schedule replaces exposures rather than adding targets. The source-position calibration counts 157,797,196 valid occurrences per cycle, excluding PAD and two-step views. In the feature formulas of Appendix A, the original query moments remain fixed, while $(m_w, s_w) = (6.626530372567584, 4.089825185605686)$, $(m_g, s_g) = (-0.022947827813332, 1.125693394339381)$, and $(m_j, s_j) = (42.775451954165270, 32.094473653804320)$. Moments use the fit exposure distribution and float64 arithmetic before float32 features. Development, ordinary additions and stress views do not enter this calibration.

Equal development-selection opportunities. Each capacity first uses pilot seed 20260923 for both arms and two AdamW recipes: A has learning rate 10^{-3} and weight decay 10^{-4}; B has learning rate 3×10^{-4} and weight decay 10^{-3}. For each arm, the lowest raw-development NLL among epochs 4/8/12 determines each recipe's pilot score; the lower recipe score wins, with ties preferring A and then the earlier epoch. Pilot fits select recipes only and are not pooled with the three formal pairs. All pilot candidates are in Table 23.

Table 23: All pilot raw-development candidates, using pilot seed 20260923. Each arm independently compares the best epoch of recipes A and B.

Width	Arm	Recipe	Epoch 4	Epoch 8	Epoch 12
128	Index	A	3.421351	3.414520	3.449970
128	Index	B	3.387232	3.432847	3.597216
128	Scan	A	3.424791	3.373214	3.416495
128	Scan	B	3.381930	3.420854	3.566551
256	Index	A	3.621170	3.571001	3.564241
256	Index	B	3.408647	3.439431	3.624316
256	Scan	A	3.637550	3.613157	3.606928
256	Scan	B	3.380096	3.397830	3.571285

The $d = 256$ recipe-A pilot favors Index at every candidate: its Index$-$Scan development differences at epochs 4/8/12 are $-0.016380/-0.042156/-0.042687$ bits/target. This is a single pilot seed, not a formal three-pair or test result. Recipe B is selected because each arm has a lower best absolute development NLL under B, independently of the sign of the between-arm difference.

Formal seeds 20260919/20260920/20260921 each complete 20 epochs and 168,000 updates from untrained initialization, with matched initialization and target order within capacity/seed, batch size 64 and FP32. Across capacities, target orders match but different tensor shapes preclude identical weight bytes. Each arm selects one endpoint by its three-seed mean raw-development NLL over epochs 4/8/12/16/20, breaking exact ties earlier. This is neither per-seed selection nor the old six-model common-endpoint rule. The actual $d = 128$ selections are Index B/epoch 4 (33,600 updates) and Scan A/epoch 8 (67,200 updates). At $d = 256$ both arms select B/epoch 4, so their actual recipes and selected update counts match. The $d = 128$ contrast is a comparison of systems given equal selection opportunities, with different final optimization recipes and update counts.

For $d = 128$, the development gap is 0.029478 bits/target at the two selected endpoints (Index B/4 versus Scan A/8). Comparing those same recipes at common epoch 4 gives 0.003424; at common epoch 8 it gives 0.078653. Both common-epoch contrasts still compare recipes B and A, so neither is a matched-recipe comparison. These descriptions show sensitivity to the development selection rule; they are not alternative test estimates and do not reselect or rescore any checkpoint on the evaluation material.

Recorded amendment and full development curves. The original $d = 128$ formal plan ended at epoch 12 with candidates 4/8/12. A recorded amendment on 2026-09-23 extended both capacity

budgets to epoch 20 and the candidate set to 4/8/12/16/20. The pilot and some early $d = 128$ development results were already visible; the amendment preceded the new causal ordinary/stress scoring. It is a documented change during execution, not a protocol preregistered in full from the start of the research. The $d = 128$ runs continued from their complete epoch-12 states; earlier checkpoint identities, losses, initialization and target orders were retained. The five-candidate rule happens to select the same endpoints as the earlier three-candidate rule, but final scoring binds the five-candidate selection. All 30 formal candidate checkpoints per capacity were retained and their archive-member hashes checked. Every formal scoring receipt binds the selected checkpoint and records unchanged weights with zero training updates.

Table 24: Complete formal development curves: three-seed means in bits/target. The fixed per-arm rules select d128 Index/Scan at epochs 4/8 and both d256 arms at epoch 4.

Width	Arm	Epoch 4	Epoch 8	Epoch 12	Epoch 16	Epoch 20
128	Index	3.402728	3.451903	3.604147	3.793945	4.017843
128	Scan	3.399305	3.373250	3.399390	3.518936	3.594738
256	Index	3.401305	3.423186	3.612641	3.958373	4.378739
256	Scan	3.378279	3.396947	3.575898	3.892777	4.270574

Table 25: All individual formal development NLLs. Seed suffixes 19/20/21 denote 20260919/20260920/20260921. No candidate or seed is dropped.

Width	Arm	Seed	Epoch 4	Epoch 8	Epoch 12	Epoch 16	Epoch 20
128	Index	19	3.389541	3.444501	3.588243	3.777771	3.994058
128	Index	20	3.414300	3.453788	3.613049	3.827587	4.047162
128	Index	21	3.404344	3.457421	3.611150	3.776475	4.012308
128	Scan	19	3.388797	3.354219	3.385507	3.494889	3.565778
128	Scan	20	3.408685	3.383738	3.402418	3.533211	3.599276
128	Scan	21	3.400432	3.381793	3.410245	3.528709	3.619161
256	Index	19	3.406281	3.423969	3.597331	3.938793	4.326494
256	Index	20	3.392578	3.421003	3.596704	3.936658	4.335607
256	Index	21	3.405056	3.424585	3.643889	3.999667	4.474117
256	Scan	19	3.377949	3.392465	3.574318	3.869322	4.235689
256	Scan	20	3.373761	3.388834	3.546405	3.875806	4.232963
256	Scan	21	3.383127	3.409542	3.606970	3.933202	4.343069

Table 24 includes all development candidates, with full seed records in Table 25. Late losses deteriorate, especially at $d = 256$; its epoch-20 Index/Scan means are 4.378739/4.270574. Three of four selected arm endpoints lie at epoch 4, the earliest candidate, so this grid does not establish whether an earlier endpoint would be better. No dropout or denser early-epoch grid was evaluated in these runs. Neither completed training budgets nor positive test gaps establish convergence or fully tuned capacity limits. For context, the old causal $d = 64$ selected per-arm development means were 3.426794/3.392741, not the six-model mean 3.409768. Both new capacities improve those corresponding per-arm values, with their changed tuning/selection procedures disclosed.

Primary ordinary prediction. Both capacities retain the original 2,662,587 targets, 2,951 records and 611 source-file blocks, including output UNK and empty-source targets. Each uses the original record weight divided by the original target count; no test-dependent filtering or reweighting is applied. The separately specified primary endpoint at each capacity is the equal-seed mean of ordinary NLL(index)−NLL(scan). All absolute seed results are in Table 26.

Table 27 collects the paired differences and distinguishes fixed-fit material intervals from descriptive variation across training pairs.

The material bootstrap uses 10,000 paired PCG64 draws, seed 20260923, resampling the same 611 original blocks for both arms and all fixed fits, and recomputing sampled original-weight mass. Linear percentile intervals are conditional on the three fitted pairs and their selected endpoints. They do not incorporate retraining, the shared pilot, development selection, or adaptive research design.

Table 26: All causal-capacity ordinary results at the frozen per-arm selections. NLL uses original weights; accuracy is descriptive.

Width	Seed	Index NLL	Scan NLL	NLL gain	Index acc. (%)	Scan acc. (%)
128	20260919	3.354981	3.317917	0.037064	52.9163	53.2337
128	20260920	3.372357	3.357491	0.014866	52.6037	52.8494
128	20260921	3.361046	3.350812	0.010234	52.5200	52.7436
128	Mean	3.362794	3.342073	0.020721	52.6800	52.9423
256	20260919	3.376969	3.346549	0.030420	52.5809	52.9684
256	20260920	3.355595	3.339145	0.016450	52.8411	52.9555
256	20260921	3.365280	3.344638	0.020641	52.6907	52.8731
256	Mean	3.365948	3.343444	0.022504	52.7042	52.9323

Table 27: **Material and training variation answer different questions.** NLL gains in bits/target. Material 95% intervals resample blocks with the three fitted pairs fixed; fit intervals are descriptive t intervals with two degrees of freedom, without selection adjustment.

	$d = 128$, 4 blocks	$d = 256$, 4 blocks
Three paired differences	.03706 / .01487 / .01023	.03042 / .01645 / .02064
Conditional material 95% interval	[.01587, .02541]	[.01972, .02500]
Across-fit SD	.01434	.00717
Descriptive fit 95% interval	[−.01491, .05635]	[.00470, .04031]

Table 28 separately reports descriptive seed uncertainty and a later two-capacity Bonferroni sensitivity. The latter uses 97.5% material intervals per size and was added after results were readable; it does not retroactively create a prespecified joint family. The three reused seed IDs do not become six independent repetitions across capacities.

Table 28: Ordinary NLL uncertainty (bits/target). Original 95% intervals resample material conditional on fixed fits; the 97.5% sensitivity and seed-t intervals are later descriptive checks. Seed-t uses $df = 2$ and does not account for shared selection.

Width	Material 95%	Material 97.5%	Seed SD	Seed-t 95%
128	[0.015867, 0.025410]	[0.015186, 0.026070]	0.014342	[-0.014905, 0.056348]
256	[0.019723, 0.024996]	[0.019306, 0.025338]	0.007169	[0.004696, 0.040312]

The $d = 128$ seed interval crosses zero, whereas its fixed-fit material interval is positive. The $d = 256$ seed interval is positive, but it still uses only three differences and an untestable normal-difference approximation. These distinctions remain part of the result. Ordinary accuracy gains are 0.262265 and 0.228111 percentage points. Both arms' ordinary NLLs improve over old causal $d = 64$; new $d = 128$ Index (3.362794) is numerically close to old $d = 64$ Scan (3.363211), and new Scan reaches 3.342073. Numerical closeness is not a test of equivalence. Increasing from $d = 128$ to $d = 256$ slightly worsens both mean NLLs (to 3.365948/3.343444). The cross-capacity comparison changes tuning and selected systems, so it supplies neither a pure capacity effect nor a monotonic scaling law or efficiency claim.

Secondary stress behavior and unfavorable cases. The previously studied stress population retains 6,493 targets, 1,069 source rows and 512 components, with target-within-row, row-within-component and equal-component weights. It is kept separate from ordinary additions. The original two-step additive interaction K_2 is secondary in these capacity experiments; no outcome-based change of primary endpoint is made. Its material intervals use 10,000 paired component draws with PCG64 seed 20260922. Table 29 retains every seed's interaction and the supplementary proportional contrast T defined in Appendix D.

Table 29: All paired-seed secondary K_2 (bits) and post-hoc proportional T (natural-log ratio difference). Mean T averages seed-level logs; it differs from taking logs after mean JS.

Width	Seed	K_2	T
128	20260919	0.006559	0.315840
128	20260920	0.005473	0.544527
128	20260921	0.003592	0.166992
128	Mean	0.005208	0.342453
256	20260919	0.014266	-0.001927
256	20260920	0.004688	0.218054
256	20260921	0.008214	0.416219
256	Mean	0.009056	0.210782

Table 30: Five-view causal-capacity means, reported as Index/Scan. NLL, entropy H and JS are in bits; accuracy and flips are percentages. All views retain the same stress targets and weights.

Width	View	NLL	Accuracy (%)	H	JS to raw	Flips (%)
128	raw	4.2914/4.2065	35.57/36.39	4.0431/3.7321	0.00000/0.00000	0.00/0.00
128	L1	4.2925/4.2063	35.36/36.19	4.0121/3.7071	0.00702/0.00458	10.57/7.59
128	R1	4.2895/4.2033	35.40/36.16	4.0452/3.7239	0.00288/0.00259	6.02/4.14
128	L2	4.3204/4.2201	35.08/36.26	4.0158/3.6816	0.01542/0.00970	15.27/10.95
128	R2	4.2884/4.2092	35.42/36.09	4.0490/3.7076	0.00488/0.00437	7.99/6.03
256	raw	4.3246/4.2647	35.07/36.01	4.0559/4.0742	0.00000/0.00000	0.00/0.00
256	L1	4.3284/4.2723	34.85/36.07	4.0197/4.0648	0.00893/0.00434	12.28/7.93
256	R1	4.3093/4.2556	35.21/35.87	4.0430/4.0826	0.00383/0.00250	6.95/4.81
256	L2	4.3726/4.2786	33.92/35.90	4.0200/4.0494	0.02154/0.00910	18.96/12.10
256	R2	4.2981/4.2549	35.35/35.96	4.0489/4.0965	0.00747/0.00409	9.93/6.57

Both mean side-specific two-step JS values decrease, but $d = 128$ seed 20260920 has worse right-side JS for Scan at both intensities: Index$-$Scan is -0.001173 for R1 and -0.000783 for R2. At $d = 256$, seed 20260919 has the largest additive K_2 but negative $T = -0.001927$; positive additive interaction does not imply the same direction under proportional scaling. The $d = 256$ descriptive seed-t 95% interval for K_2 is [-0.002978, 0.021090], whereas the $d = 128$ interval is [0.001479, 0.008936]. They have the same three-fit limitations as the NLL seed intervals and do not replace the primary NLL results. Raw stress entropy decreases from 4.043145 to 3.732134 bits at $d = 128$ but increases from 4.055924 to 4.074204 at $d = 256$. These measurements do not isolate temperature or calibration effects. Full prediction distributions were not saved for these capacity scores, so the present analysis cannot independently recompute each JS from logits or perform temperature matching. Argmax flips are not errors; the stress population's enriched internal-run eligibility also prevents interpreting flip rates as prevalence on ordinary material or as free-generation quality.

Earlier larger bidirectional configuration. An earlier $d = 128$, four-layer bidirectional W comparison (1,181,387 parameters) used recipe A and the original six-model common-epoch rule, selecting epoch 8 from the five candidates. Its ordinary NLLs were 3.413978/3.378184, with paired gains 0.041410/0.034947/0.031024 and mean 0.035794. Both arms improve over the smaller bidirectional configuration, but neither surpasses its corresponding old causal $d = 64$ arm. This boundary is retained alongside the later causal checks. The larger bidirectional study's *original* primary was $K_2 = 0.006178$ with material 95% interval [0.005071, 0.007347], paired values 0.008174/0.002657/0.007703, and later seed-t interval [-0.001420, 0.013776]. Its ordinary NLL was originally secondary and is not relabeled as a preregistered primary. K_2 is smaller than in the old small bidirectional model, and one seed's one-step NLL interaction $\Gamma_1 = -0.006531$ is negative. Selected and fixed-epoch-8 scoring are identical and count once. The complete development table is retained below.

Reproducibility scope. Independent reaggregation of saved scores reproduces ordinary weighted losses and stress aggregates, with maximum absolute discrepancy 1.95×10^{-14} across 126 scalar comparisons; ordinary bootstrap replicas also agree elementwise with a separate full-draw imple-

Table 31: Complete earlier bidirectional d128 development candidates. The six-model mean selects epoch 8; fixed-epoch-8 and selected scores are identical.

Epoch	Index 19	Scan 19	Index 20	Scan 20	Index 21	Scan 21	Mean
4	3.463546	3.425626	3.460624	3.441676	3.477009	3.445995	3.452413
8	3.435521	3.393802	3.442554	3.401026	3.441968	3.414587	3.421576
12	3.473001	3.438782	3.496193	3.469531	3.523671	3.446074	3.474542
16	3.620448	3.528147	3.634136	3.564305	3.628297	3.579541	3.592479
20	3.706298	3.598016	3.713628	3.633315	3.730394	3.647479	3.671522

mentation. No new model forward pass, training or pixel rendering was performed for that verification. Complete new statistical outputs, candidate development records and protocol audits are retained in `supplementary/new-results/causal-capacity/` and `supplementary/new-results/strong-bidirectional/`. Their aggregate files supply numerical replay without redistributing natural source texts or checkpoint bodies. The common single-font, reused-material and three-paired-fit scope remains unchanged.

D WIDTH-KNOWN MODELS WITH CAUSAL SOURCE ENCODING

This section records the initial causal-source W experiment. The bidirectional source-mask follow-up is reported separately in Appendix E.

Matched information and capacity. W-index and W-scan tested whether precomputed cumulative coordinates reduce sensitivity to whitespace recoding when both models already know each source character's true advance. Both retained the full previous row, character IDs, sinusoidal indices, up to 128 characters of target history, and the same legal target query (c, q) from Appendix G. Both added a bias-free $2 \to 64$ source projection and had 348,811 parameters. The first scalar was the standardized true local advance, looked up using the original Unicode character even when its model ID was UNK. The second was a moment-matched character index in W-index and a standardized cumulative coordinate in W-scan. Specifically, the latter used $g_j = (\sum_{t<j} w_t + 0.5 - \mu_q)/\sigma_q$ with positive, uncentered advances w_t; summing standardized widths would give a different feature. The index scalar matched the mean and variance of g_j over the actual fit-view source position occurrences. Width moments used those same occurrences; padding was excluded. The target query was unchanged, so the experiment does not test learning to accumulate the current target-row prefix.

Matched training and endpoint selection. Six fresh fits used seeds 20260919, 20260920 and 20260921, with identical initial weights, target order and view schedule within each pair. The two-layer $d = 64$, four-head, FF=256 architecture, shared causal source/history encoder, full-source cross-attention, output UNK mass and probability floor were retained. All 537,584 targets from the 638 fit records participated. For targets with at least five U+3000 spaces in each eligible internal run, a hash-phased four-epoch cycle presented R, R, L_1, R_1; other targets always received R. Thus augmentation replaced an exposure rather than multiplying examples. Each fit completed 20 epochs/168,000 updates with AdamW (learning rate 0.001, weight decay 0.0001), batch size 64, FP32 and no dropout. L_2/R_2 entered neither training, scalar calibration, nor development selection. The six-model mean raw-development NLL selected one common endpoint from epochs 4/8/12/16/20, with exact ties going earlier. Their mean losses were 3.432792/3.409768/3.417256/3.468035/3.490765 bits, selecting epoch 8 (67,200 updates). Each candidate completed whole exposure cycles. This is a new width-known comparison, not canonical-input training or a one-variable comparison with the earlier checkpoints.

Input-frozen five-view population. Each eligible target had an internal whitespace run on either side of its unchanged query q, with visible ink beyond both ends of each run. A left run ended at or before q; a right run began at or after q. Each selected run contained at least ten U+3000 spaces. In the fixed MS PGothic 16-pixel metric, U+3000 and U+0020 had advances 11 and 5 pixels. Intensity k removed the first $5k$ U+3000 characters in one run and inserted $11k$ U+0020 characters at the first removal, adding $6k$ characters while preserving the other characters and the $55k$-pixel total advance.

Mixed runs need not contain a contiguous $55k$-pixel substring. A fixed input-identity hash chose each run pair. The same target then yielded R, L_1, R_1, L_2, R_2; each L_k/R_k pair had equal total character counts, space composition, advances and full-row pixels. Targets, histories, queries and labels were unchanged, and every source feature was rebuilt by view. Full-canvas checks covered all 25,972 recoded target–view comparisons; all matched, and a one-pixel shift was detected for every target bundle.

The input-only candidate frame contained 120,377 components and 133,893 records; 16,022 components, 18,161 records, 53,277 source rows and 474,332 targets met the two-step eligibility rule. Fixed hashes selected 512 components, one record per component, at most four source rows per record and at most 16 targets per row: 512 records, 1,069 rows and 6,493 targets. Both intensities used exactly this population, fixed before new model outcomes were read. Components received equal weight, source rows equal weight within component, and targets equal weight within row; the three fitted seeds were then averaged equally. Views are paired measurements, not additional sampling units. This enriched internal-whitespace population is separate from the old additions population and its source-block weights.

All-run canonicalization. Extending the earlier leading-space rule to every maximal U+0020/U+3000 run chooses nonnegative (a, b) minimizing $(a + b, a, b)$ lexicographically subject to $5a + 11b = W$, then emits b ideographic and a ordinary spaces. Each rewrite preserves its run's advance W, so all five views map to the same source string. An audit of the existing 6,493 heldout targets verified exact string equality, idempotence, and preservation of total advance and nonspace-character pen origins against the frozen Unicode manifests and NPZ arrays; leading-only normalization did not merge the five views for any target. No training, model forward passes or new rasterization were performed. The original W stress study retains distinct, uncanonicalized views as a probe of representation behavior. This audit alone establishes preprocessing identity and supplies no NLL result. The separately fitted all-run experiment in Appendix A now tests ordinary prediction after that preprocessing. Detailed receipts are in `supplementary/all-space-identity/`.

Historical-use boundary. The design followed earlier findings. Before reading candidate raw content, a usage ledger bound old fit/dev/evaluation, the 713-component development study, the 64-component gap study, the 4,096-component baseline pool and eight design examples, alongside registered exclusions including sealed P1, an excluded sealed held-out pool that contributes no results to this paper. The allowlist had no overlap with these bound uses or their component and whole-record normalized/whitespace relations. This is a finite recorded-use check: representative source-occurrence links do not exhaust alternative duplicate provenance, authors, source files or local templates. It therefore supports an input-frozen new stress population within the recorded exclusions, not a claim of complete historical independence or public preregistration.

Estimands and conditional uncertainty. For arm a and side s, let $J_{a,s,k} = \mathrm{JS}[p_a(\cdot \mid R), p_a(\cdot \mid s_k)]$ in bits, with the fixed output support, floor and UNK treatment. Define

$$K_k = (J_{\mathrm{index},L,k} - J_{\mathrm{index},R,k}) - (J_{\mathrm{scan},L,k} - J_{\mathrm{scan},R,k}).$$

Only K_2 was primary. Its estimate was 0.00713959 bits, with paired component percentile 95% interval [0.00606186, 0.00826940], using 10,000 PCG64 draws (seed 20260922). Each draw retained all arms, seeds, views, rows and targets together. The interval conditions on the three fitted pairs and common selected endpoint; it does not resample training or the research process. Secondary K_1, $K_2 - K_1$, side-specific JS differences, and $\Gamma_k = (\mathrm{NLL}_{\mathrm{index},L_k} - \mathrm{NLL}_{\mathrm{index},R_k}) - (\mathrm{NLL}_{\mathrm{scan},L_k} - \mathrm{NLL}_{\mathrm{scan},R_k})$ are descriptive.

Joint reading of the outcomes. K_2 was positive in every fitted pair (Table 32). The mean JS reductions were 0.00294105/0.00096422 bits on the left/right at $k = 1$ and 0.00934812/0.00220852 at $k = 2$. Thus the mean interaction included improvement on both sides, with the larger improvement before the query. The second seed nevertheless had slightly *higher* W-scan right-side JS: 0.00464946 versus 0.00430203 at $k = 1$, and 0.00740190 versus 0.00673286 at $k = 2$; its right-side flip rates also increased. Γ_1 was slightly negative for the first seed. These deviations remain in Tables 32 and 33. The seed-averaged W-scan results had lower absolute NLL and higher accuracy in all five views, with comparable entropy. This joint pattern does not support explaining the lower mean JS solely by degraded, nearly constant predictions.

Table 32: All paired-seed interactions (bits). Seeds 1/2/3 denote 20260919/20260920/20260921. SD is the sample SD across three fits; the range is descriptive, not a confidence interval.

Quantity	Seed 1	Seed 2	Seed 3	Mean	SD	Range
K_1	0.000735	0.003044	0.002152	0.001977	0.001164	[0.000735, 0.003044]
K_2	0.004859	0.010021	0.006538	0.007140	0.002633	[0.004859, 0.010021]
$K_2 - K_1$	0.004125	0.006977	0.004386	0.005163	0.001577	[0.004125, 0.006977]
Γ_1	-0.000220	0.001057	0.014734	0.005190	0.008290	[-0.000220, 0.014734]
Γ_2	0.013857	0.026349	0.028707	0.022971	0.007980	[0.013857, 0.028707]

Table 33: Absolute five-view metrics for the stress population. Each cell reports W-index/W-scan, with the same hierarchical weights in every row. NLL, entropy H and JS are in bits; accuracy and argmax flip rate are percentages. JS and flips compare each view with its own arm's raw prediction. Mean denotes equal averaging over all three seeds; no model or target is excluded by prediction correctness.

Seed	View	NLL	Accuracy (%)	H	JS to R	Flips (%)
Mean	R	4.2830/4.2417	36.14/36.72	3.8495/3.8539	0.00000/0.00000	0.00/0.00
	L_1	4.2939/4.2489	35.99/36.79	3.8157/3.8390	0.01104/0.00810	12.00/10.42
	R_1	4.2883/4.2485	35.83/36.70	3.8479/3.8483	0.00466/0.00369	5.99/5.84
	L_2	4.3241/4.2582	35.58/36.34	3.8034/3.8347	0.02476/0.01541	18.37/14.05
	R_2	4.2977/4.2548	35.64/36.48	3.8431/3.8464	0.00809/0.00588	8.48/7.35
1	R	4.2855/4.2318	36.09/36.88	3.8461/3.8467	0.00000/0.00000	0.00/0.00
	L_1	4.2950/4.2470	35.43/36.54	3.8148/3.8368	0.01181/0.00898	12.40/11.32
	R_1	4.2797/4.2315	35.52/36.65	3.8157/3.8381	0.00502/0.00292	6.38/4.73
	L_2	4.3209/4.2539	35.54/36.29	3.7881/3.8276	0.02350/0.01494	17.55/14.12
	R_2	4.2862/4.2330	35.50/36.65	3.7990/3.8384	0.00815/0.00445	8.61/6.21
2	R	4.2978/4.2707	35.87/36.90	3.7871/3.7838	0.00000/0.00000	0.00/0.00
	L_1	4.3123/4.2843	36.18/37.02	3.7519/3.7683	0.01043/0.00774	11.09/9.39
	R_1	4.3120/4.2850	35.44/36.75	3.7727/3.7632	0.00430/0.00465	5.37/6.05
	L_2	4.3457/4.2919	35.26/36.12	3.7676/3.7708	0.02487/0.01552	17.20/13.55
	R_2	4.3254/4.2980	35.39/36.38	3.7739/3.7599	0.00673/0.00740	6.80/7.77
3	R	4.2658/4.2226	36.47/36.38	3.9154/3.9313	0.00000/0.00000	0.00/0.00
	L_1	4.2743/4.2154	36.36/36.79	3.8805/3.9120	0.01088/0.00758	12.49/10.56
	R_1	4.2731/4.2289	36.52/36.71	3.9554/3.9436	0.00466/0.00351	6.21/6.75
	L_2	4.3056/4.2289	35.94/36.61	3.8545/3.9057	0.02590/0.01577	20.35/14.49
	R_2	4.2815/4.2334	36.04/36.41	3.9562/3.9408	0.00937/0.00578	10.02/8.08

A separate raw reference on the *previously used* additions population had W-index/W-scan NLLs 3.39298943/3.36321088 and a descriptive gap of 0.02977855 bits. Its three paired gaps were 0.05429592, 0.01939250 and 0.01564725. The archive name `raw_fresh_reference` does not make this population newly held out; it is neither pooled with the stress test nor a second primary confirmation. The stress result supports reduced extra sensitivity to query-preceding recoding under this fixed font and training scheme. Recoding also changes token density, sinusoidal positions and causal source prefixes, so it does not uniquely identify arithmetic computation, unseen-length generalization or free-generation layout quality.

Scale of the directional contrast. A post-hoc description of the same saved scores complements the additive K_2. At intensity two, the left/right relative JS reductions computed from the equal-seed mean JS are 37.75%/27.32%. Define the dimensionless contrast $T = \ln(J_{\mathrm{index},L,2}/J_{\mathrm{index},R,2}) - \ln(J_{\mathrm{scan},L,2}/J_{\mathrm{scan},R,2})$. Its three seed values are -0.15253, 0.56630 and 0.01273; applying the contrast to mean JS gives 0.15504, distinct from the mean seed-level contrast 0.14217. All inputs are positive and no epsilon is used. Thus the three positive additive interactions do not imply a common direction on the proportional scale. In particular, uniform contraction $J_{\mathrm{scan},s,2} = cJ_{\mathrm{index},s,2}$ gives $K_2 = (1-c)(J_{\mathrm{index},L,2} - J_{\mathrm{index},R,2}) > 0$ for $0 < c < 1$ and larger index JS on the left, while $T = 0$. Comparable predictive entropy does not exclude this overall contraction. These descriptive diagnostics retain the primary endpoint and add no significance tests; the script and complete values are in `supplementary/k2-scale/`.

E BIDIRECTIONAL SOURCE ENCODING AT FIXED CAPACITY

Architectural intervention. This follow-up changes the source self-attention mask of both W arms from causal to bidirectional. Target-history encoding remains causal, and the prediction query retains access to every valid source position. Local Unicode widths, source-position scalars, target inputs, vocabulary, calibration and 348,811-parameter capacity are unchanged. Source and history still reuse encoder weights within each model. The seeds 20260919/20/21 reuse the original causal study's untrained initial states, RNG states, target orders and R, R, L_1, R_1 exposure schedules. These are new fits under a different mask, not additional independent seed draws. The fit budget is again 20 epochs and 168,000 updates per arm, with batch size 64, learning rate 0.001, weight decay 0.0001, FP32, TF32 disabled and no dropout.

Selection and materials. The six-model mean raw-development NLL at epochs 4/8/12/16/20 is 3.470825/3.453527/3.463780/3.527722/3.550436 bits, selecting epoch 8 (67,200 updates). The selected-endpoint and separately executed fixed-epoch-8 score files are byte-identical; we count them once. The ordinary reference uses the same 2,951 additions records, 2,662,587 targets and 611 blocks. The stress analysis reuses the same 512 components, 1,069 source rows, 6,493 targets and five views from Appendix D. The source-mask protocol was fixed before these new fits, after the causal results were known. Neither evaluation population is a new blind confirmation.

Prediction and directional response. Table 34 separates the ordinary prediction difference from the interaction on matched stress targets. Each is positive in every pair. The K_2 interval is [0.007628, 0.010232] bits, from 10,000 paired-component percentile bootstrap draws (PCG64 seed 20260922, linear quantiles), averaging rather than resampling the three seeds. Its inferential scope is conditional on these fits and the common selected endpoint. The across-pair sample SD of K_2 is 0.002554 bits; it describes these three paired differences and does not turn the material interval into a retraining interval.

Ordinary mean accuracies are 51.2249% for W-index and 51.7171% for W-scan. On the stress population, mean JS reductions for L1/R1/L2/R2 are 0.005137/0.002163/0.013504/0.004578 bits, and $K_2 > K_1$ in each pair. Table 35 gives absolute raw and two-step endpoints. The saved replay includes all five views, entropy and each seed. Lower seed-averaged sensitivity accompanies lower NLL and higher accuracy on all five views. This is a statement about the means: for seed 20260919, scan's R1/R2 accuracies are 0.139/0.194 percentage points below index's. Raw mean entropy is 3.952610/3.888281 bits (index/scan); entropy differences remain relevant to interpreting distributional sensitivity.

Table 34: Bidirectional source results by paired initialization. NLL gain is W-index minus W-scan on ordinary additions. K_1 and K_2 use the stress population and are reported in bits; only K_2 has a primary interval.

Seed	Ordinary NLL gain	K_1	K_2
20260919	0.045565	0.002596	0.005998
20260920	0.040881	0.002523	0.010082
20260921	0.029690	0.003806	0.010695
Mean	0.038712	0.002975	0.008925

Table 35: Bidirectional width-known models on matched stress targets. NLL is bits/target, JS is bits, and accuracy/flips are percentages. All cells average three fixed fits. Flips compare a view with its raw prediction and need not represent new errors.

Model	View	NLL	Accuracy	JS vs raw	Flips vs raw
W-index	Raw	4.374646	34.274	—	—
	L2	4.455298	32.725	0.029473	21.513
	R2	4.355560	34.707	0.013791	14.791
W-scan	Raw	4.309566	35.262	—	—
	L2	4.340677	34.465	0.015969	14.087
	R2	4.315691	34.928	0.009213	10.519

Interpretation. Persistence under the new mask shows that the positive within-architecture prediction differences and directional interaction do not require causal source encoding. The bidirectional fits have higher absolute NLL than the causal fits on the ordinary reference; this is a mask intervention, not a performance improvement from greater capacity. We do not use the two point estimates of K_2 as a test of a difference between architectures. As in the causal study, additive interactions can reflect a general reduction in response magnitude (Appendix D); they do not uniquely identify prefix-sum computation inside the network. Full internal-run canonicalization was not used in these fits.

F Earlier full-row and leading-space results

These earlier F-series comparisons establish the historical complete-row and leading-space evidence. Their original estimates, interval families and cross-reference labels are retained here; the all-run W experiment in Appendix A is a separate joint control. Model and material definitions are in Appendix G, with training-control methods in Appendix I.

F.1 Geometry helps with complete source access

The raw-trained models test whether geometry improves prediction when both models can access the complete source row (Table 36). F-geom improves over F-index by 0.0280 bits/target on the studied evaluation. Freezing those checkpoints, vocabulary, and coordinate calibration retains a 0.0322-bit advantage on screened additions. Every paired seed favors geometry in both populations. F-char is included as the zero-scalar reference; the active index accounts for part, but not all, of its gap to geometry.

Entirely known sources carry 98.03% of old-evaluation and 96.82% of additions weight, retaining conditional gains of 0.0282 and 0.0320 bits/target. Geometry helps where source IDs preserve the characters needed to reconstruct positions (Appendix H.3).

F.2 Geometry after leading-space normalization

Let C map the specified equivalent leading-space encodings to a common source string. For original row R and its recoding T, $C(R) = C(T)$; unchanged target inputs then give identical deterministic

Table 36: Base full-row models and their frozen evaluation on additions (bits/target). The populations use their own groups and weights. Each difference has a separate 95% interval; these intervals are not a joint family.

Model / difference	Studied evaluation	Screened additions
F-char (zero extra scalar)	3.578401	3.405151
F-index (matched active scalar)	3.568899	3.396202
F-geom (cumulative advance)	3.540872	3.363959
F-index − F-geom	**0.028027**	**0.032244**
95% interval	[0.022437, 0.032599]	[0.028230, 0.036233]

predictions. Both arms apply this leading-space canonicalizer during *training, development, and evaluation*. Their active scalar channels encode either fit-matched canonical-row indices or true cumulative advances. Both pipelines therefore use font knowledge, through preprocessing alone or preprocessing plus explicit source positions.

Geometry lowers NLL from **3.400790 to 3.360967 bits/target**, a difference of **0.039823** with a 97.5% interval of **[0.036257, 0.043532]**. All three fitted pairs favor geometry: 0.05399, 0.03822, and 0.02727 bits/target. The interval belongs to the original two-contrast family, which also includes a pseudo-coordinate comparison (Appendix I). A descriptive check of sources whose tokens actually change under leading-space canonicalization retains a 0.0344-bit gain on 23.34% of the population weight. The predictive benefit therefore survives removal of the tested leading-space redundancy, including on sources affected by that removal.

Two complementary controls are retained in Appendix I. Real coordinates outperform cumulative pseudo-coordinates with exact whitespace widths by 0.0328 bits/target, within the shared real-target metric. Equal one-step leading-space augmentation retains a 0.0418-bit prediction advantage and lower two-step JS sensitivity. These establish additional within-study comparisons; the W studies in Appendices D and E directly share local width information and test where equal-sized rewrites affect prediction.

G Full-row models, inputs, and fitting

G.1 Inherited pools and causal inputs

The fit/development/evaluation pools have 638/654/632 records, 581/581/573 components, and 537,584/549,663/497,645 targets. Components are disjoint. The pools inherit an input-only common-support restriction: an originally sampled component is retained when every originally sampled record has defined positive advances under MS PGothic 16px and Osaka 4.2 Bitmap 12px for its complete loader text, including padding. This checks sampled records, not all unsampled component members. It defines support; it is not an Osaka full-row experiment.

The loader decodes entities, right-strips lines, and removes boundary blank lines. The full-row builder reconstructs actual previous rows without extending them to drawing width or further trimming. Existing source spaces remain. Eligible targets are non-TAB/non-Unicode-Zs characters on a row with at least two such characters, with an actual previous panel row and another panel row s satisfying $|s - r| > 1$. There is no random floor-row draw in the neural models. Source lengths reach 228/200/196 in fit/development/evaluation.

History is $x_{r,\max(0,c-128):c}$ with masked left PAD. Both target q_{MS} and source g_j follow Equation 1, using the MS PGothic face-2, 16px table. Target coordinates use fit means $(40.9426322212, 281.0954529897)$ and standard deviations $(26.9017378079, 164.0111223765)$ for (c, q_{MS}); source geometry uses the target-q scale. The complete prefix determines q, even beyond the history window; neither target width nor future suffix is used.

G.2 Architecture, vocabulary, and optimizer

Each base F-series model has 348,747 parameters. Shared 64-dimensional embeddings feed history and source streams that reuse two pre-layer-normalized causal blocks: four heads, $64 \to 256 \to 64$ GELU feedforward modules, zero dropout. History has learned positions and an appended QUERY; source has sinusoidal indices and a bias-free $1 \to 64$ scalar projection. QUERY cross-attends with four heads to all valid source states. PAD keys are masked; an empty source enables a learned NULL key/value. Retrieved and residual query states are normalized, concatenated with two target coordinates, and passed to a $66 \to 128 \to 1099$ tanh head. Coordinates also enter QUERY's projection.

The fit-only vocabulary includes complete fit sources, adding U+6E80 and U+773C to yield 1,135 input IDs including PAD. There are 1,099 known output classes plus an unknown outcome. With $\epsilon = 10^{-4}$, known-target probability is $(1-\epsilon)p_y + \epsilon/1100$; unknown targets receive $\epsilon/1100$. The 629 old-evaluation unknown targets and 681 empty-source targets stay in the score. Coordinate construction uses original characters before vocabulary mapping.

Seeds 20260913/20260914/20260915 pair initialization and shuffled target order. Arms train separately for 20 epochs (168,000 updates), using FP32, batch 64, and AdamW with learning rate 0.001 and weight decay 0.0001. Training minimizes unweighted minibatch mean smoothed NLL; population weights apply to development and evaluation. F-char zeros only the extra scalar, leaving its 64 projection weights allocated but inactive.

G.3 Matched active index and checkpoint selection

For normalized geometry $\widetilde{g}_j = (g_j - \mu_q)/\sigma_q$, F-index supplies

$$h_j = m_g + \frac{s_g}{s_j}(j - m_j) = 0.035111572757626094j - 1.5227212790330815.$$

Moments are uniform over 39,385,873 valid source-position occurrences across 537,584 training targets; repeated reads of a source row repeat its positions. Population standard deviations use divisor N. Stored float64 moments are

$$(m_g, s_g) = (-0.023082261040306872, 1.125635578744892),$$
$$(m_j, s_j) = (42.71067628740894, 32.058819652287106).$$

The affine scalar is computed in float64 then cast to float32. Invalid positions stay zero and masked. Calibration uses no development or evaluation data. Canonical and other training extensions apply their own corresponding fit calibration as specified in Appendix I.

The original F-char/F-geom comparison selects one common checkpoint by the mean development NLL of its six models. The later F-index extension trains three index trajectories and reuses the five candidate checkpoints of each original geometry trajectory. It performs a separate six-model selection using F-index/F-geom only, never a nine-model selection. Initial-parameter and every-epoch target-order hashes match within pairs. Candidates are epochs 1/2/5/10/20, with earlier exact ties; both comparisons select epoch 10. All three arm-averaged curves are lowest at that candidate and rise at 20.

Table 37: Full-row development NLL (bits/target). Each arm averages three seeds. The original F-char/F-geom comparison and the later F-index/F-geom control use separate six-model means; both select epoch 10 from the same five candidates, with earlier exact ties. F-char does not enter the later selection; no nine-model mean is used.

Epoch	F-char	F-index	F-geom	Original six-model mean	Index control six-model mean
1	3.604114	3.589044	3.587055	3.595585	3.588050
2	3.514320	3.509905	3.501398	3.507859	3.505651
5	3.443347	3.433553	3.414837	3.429092	3.424195
10 (selected)	3.436759	3.424737	3.390844	3.413801	3.407790
20	3.524897	3.509665	3.462663	3.493780	3.486164

Development and evaluation use Equation 3. For each original comparison, 10,000 paired component-bootstrap draws (PCG64 seed 4461859123330840698) use linear 0.025/0.975 quantiles. F-char/F-geom and F-index/F-geom are separate one-contrast families. Table 39 reports every fixed pair. The F-index old-evaluation difference is 0.028026619, interval [0.022436639,0.032598737]. The original zero-channel difference is 0.037529, interval [0.032886,0.041623]. Neither interval quantifies retraining or checkpoint-selection variation.

H SCREENED ADDITIONS AND FROZEN EVALUATION

The current source-version documentation is identified explicitly.

H.1 DATA PROVENANCE AND SOURCE VERSIONS

HukuTemp compiles the community-maintained Yaruo-thread MLT Collection, whose public collection thread links its distribution channels and the AAMZ Viewer (Yaruo-thread MLT Collection, 2026). The inherited pools use the v32.0 derivative dataset `1anon8anon1/yaruo-mlt-aa`, fixed at revision `a95fa076383f788b38dc76ecd5f6900313d6e3cd` (1anon8anon1, 2026). The additions instead compare archived v32.1 and v32.2 ZIP releases. The three archive READMEs date v32.0 to June 27, v32.1 to July 26, and v32.2 to September 6, 2026. They share a statement permitting redistribution and reposting of "this information." This statement does not specify a standard data license or establish rights for each community artwork or font. The current package's `supplementary/new-results/data_provenance.json` records the public source links, fixed snapshot revision, and archived ZIP/README hashes; it is version documentation, not a complete natural-data reconstruction entry point. The ZIP releases and the derived dataset snapshot are distinct containers, not asserted byte-identical copies.

H.2 SCREENING, GROUPING, AND COMPLETE DENOMINATORS

The HukuTemp v32.1-to-v32.2 difference gives 3,163 distinct normalized-text candidates: 2,951 retained, 77 quarantined for unresolved recorded relations, and 135 out of scope (134 shorter than 15 rows and one missing a table width). The retained 2,960 occurrences span 641 source paths; one predetermined representative per normalized hash supplies each record. Performance and OOV rates do not select records.

Screening against bound v32.0/v32.1 references checks normalized-text identity, identity after whitespace removal, and Jaccard similarity at least 0.5 between sets of whitespace-collapsed lines of length at least eight. Short/narrow references remain eligible for matching. Known-use component closures, registered external inputs, and replacement flags add checks; a candidate linked to a known training component is quarantined. Direct record-ID and normalized-hash intersections with the original fit/development/ evaluation pools are zero. The audit does not certify absence of crops, mirrors, semantic derivatives, unregistered use, or relations outside the references. Of 641 paths, 628 modify existing paths; this is same-tradition material, not an independent-author or second-font sample.

Joining 2,520 detected relation clusters across shared files yields 611 blocks. They group detected dependence, not certified independent authors. All 2,951 records and 2,662,587 eligible targets stay in the estimand, including 5,431 unknown targets and 1,915 empty sources. Parsing and target eligibility match the full-row rules above; target order is normalized SHA, row, column. For stratum S, original target weights are $w_{i,t} = 1/(Bn_bT_i)$, mass $p_S = \sum_S w_{i,t}$, contribution $B_S = \sum_S w_{i,t}d_{i,t}$ for seed-averaged difference d, and conditional difference B_S/p_S. Complementary contributions reconstruct the overall contrast without changing denominators.

H.3 FROZEN CHECKPOINTS AND SOURCE COVERAGE

The six original full-row checkpoints are scored with their epoch-10 weights, original vocabulary and coordinate calibration. Characters are re-encoded with the full-row vocabulary, not local-model IDs. The subsequent F-index transfer scores only its three fixed checkpoints, reusing the geometry exports after checking checkpoint, input/target-order, and prediction identities. There is no retraining, checkpoint reselection, vocabulary expansion, or scale refitting.

Stored float32 log probabilities are reduced in float64. Original F-char transfer uses 10,000 paired block draws with PCG64 seed 13656097254779995238; F-index transfer uses seed 2188592941351691906. Both use linear 0.025/0.975 quantiles, in separate families. F-index transfer gives 0.032243551, interval [0.028230273,0.036233436]; F-char transfer gives 0.041192607, interval [0.036024,0.045608]. Population weights differ from old evaluation, so larger point estimates do not establish growth of an effect.

The neural designs use previously examined material. Base full-row fitting follows earlier local-model studies; additions screening and local-transfer rules precede local-transfer scores, with earlier findings known. Full-row transfer follows those results, and F-index transfer follows original full-row transfer. The designs precede their respective new scores but are not independent confirmations blind to prior material results. Inference is conditional on fitted models, selected endpoints, screened frame and grouping, and excludes fitting, screening, selection and research-adaptation uncertainty.

Table 38 gives the complete known/UNK and empty/nonempty partitions for the matched-index comparison itself. Empty sources count as known. No subgroup intervals are introduced. The earlier F-char coverage and local H/C/R transfer are preserved in the separately archived legacy `previous-evidence-report.pdf`, which is not included in the current compact package; they are distinct comparisons.

Table 38: Descriptive F-index−F-geom coverage at frozen epoch 10. Original estimand mass p_S differs from target fraction; B_S retains full-record denominators and B_S/p_S is conditional (bits/target). Each adjacent pair partitions its population. No subgroup intervals are added.

Material	Source stratum	Targets	p_S (%)	B_S	B_S/p_S
Original	All IDs known	487,949	98.027247	0.027635790	0.028191947
	Contains UNK	9,696	1.972753	0.000390829	0.019811361
Original	Empty	681	0.121012	−0.000298295	−0.246500314
	Nonempty	496,964	99.878988	0.028324914	0.028359232
Additions	All IDs known	2,605,872	96.820064	0.030956782	0.031973519
	Contains UNK	56,715	3.179936	0.001286769	0.040465264
Additions	Empty	1,915	0.164654	0.000153432	0.093184796
	Nonempty	2,660,672	99.835346	0.032090119	0.032143044

Table 39: Geometry advantages for every fixed model pair, in bits/target. Within each comparison, the same epoch-10 checkpoints appear for both populations. Seed rows are descriptive; the two populations have distinct weights and primary intervals.

	F-char − F-geom		F-index − F-geom	
Paired seed	Old	Additions	Old	Additions
20260913	0.046064	0.054243	0.044436	0.044423
20260914	0.031530	0.040125	0.022897	0.028325
20260915	0.034993	0.029210	0.016747	0.023982
Mean	0.037529	0.041193	0.028027	0.032244

I TRAINING CONTROLS AND ENDPOINT SELECTION

These follow-up studies were designed after the earlier raw, recoding, and reduced-data results. Their training definitions preceded the new runs, and additions scoring followed endpoint selection. Each uses the inherited fit/development pools, the 348,747-parameter full-row interface, and seeds 20260913–20260915. Within a study and seed, arms share initialization tensors and target order. Vocabulary, target history, real target coordinates, smoothing, complete source access, and the active scalar projection remain fixed.

Fitting and selection. Training uses FP32, zero dropout, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 64 with the original 48-target tail batch, and 20 epochs (168,000

updates). Each study independently selects one common epoch by its six-model mean development NLL among epochs 1, 2, 5, 10, and 20; ties take the earliest. Canonical development sources use leading-space normalization, pseudo-coordinate sources retain raw text with their respective channels, and augmented models use raw development sources. All select epoch 10 (84,000 updates). Table 40 gives every candidate mean; the replay bundle retains every model's candidate losses and selection records.

Table 40: Complete development selection curves: six-model mean NLL in bits/target. Endpoints are selected separately for each study.

Epoch	Canonical	Pseudo-coordinate	Augmentation
1	3.587523	3.589559	3.589652
2	3.507076	3.508914	3.505574
5	3.425304	3.424229	3.419792
10	3.410304	3.412549	3.403445
20	3.484829	3.485202	3.463012

Leading-space canonical training. Both arms apply the leading-space canonicalizer in Appendix J throughout fitting, development, and evaluation. Index moments are fitted on valid canonical source positions, counting each position once per fit target accessing its row. On additions, canonicalization changes 28,384 source rows accessed by 749,766 targets. Source-bank tokens change from 4,938,406 to 4,938,286; most changes reorder spaces. The primary comparison includes all 2,662,587 targets and estimates the geometry increment after this preprocessing.

Cumulative pseudo-coordinates. Both arms read raw source text and real target coordinates. A fixed width permutation (salt 20260917) preserves whitespace advances. Nonspace Unicode code points are ordered by descending fit source-occurrence frequency, with code-point ties, then split into at most eight nearly equal-cardinality strata. Frequencies count each source position once per accessing fit target. Within each stratum, donors sorted by SHA256 of `20260917|stratum|k|codepoint` assign their positive widths to recipients in frequency order. Unseen nonspace code points use a deterministic fallback hash into the sorted fit-width multiset, before tokenization to UNK. Pseudo positions accumulate these widths with the shared 0.5 anchor. An affine transform matches the real channel's fit mean and standard deviation over valid source-position occurrences. The comparator thus retains positive cumulative advances, content dependence, monotonicity, matched moments, and an active projection. Because target coordinates remain real, the comparison includes source–target metric consistency.

Equal one-step augmentation. For every eligible fit source, each adjacent epoch pair contains one raw and one expanded view. A fixed hash of the namespace, paired seed, epoch pair, and source identity sets their order independently of target order. All targets accessing a row share its epoch view, and both arms share the masks: ten raw and ten expanded epochs per eligible row. Ineligible rows remain raw. Index moments match this fit mixture; geometry retains the true-advance scale. One expansion replaces five ideographic spaces with eleven ordinary spaces in the leading blank prefix, preserving 55px and adding six characters. Two-step expansion is excluded from fitting, calibration, and endpoint selection. Its input-defined support and behavioral results appear in Appendix J.

Primary results and inference. The canonical and pseudo NLL contrasts form one two-comparison family. The augmented full-population raw NLL contrast D_{aug} and two-step JS contrast S_{aug} form another. Within each family, three fixed seeds are averaged before 10,000 shared paired source-file-block draws (PCG64 seed 20260918). Linear 0.0125/0.9875 quantiles give 97.5% intervals and nominal 95% Bonferroni coverage within that family. All 611 original blocks remain; the JS denominator is eligible original mass in each draw, including 56 blocks with zero two-step mass. These intervals condition on fitted models and selected endpoints. Absolute NLLs and intervals are in Table 42; all paired-seed contrasts are below.

The studies reuse the same screened additions. Comparisons between their absolute losses do not isolate a training-strategy effect: fits, calibration distributions, and execution backends differ.

Table 41: All fitted-seed contrasts. NLL differences use bits/target; S_{aug} is a JS difference in bits. Positive values favor real geometry.

Contrast	Seed 20260913	Seed 20260914	Seed 20260915	Mean
Canonical index − geometry	0.053987	0.038216	0.027268	0.039823
Pseudo − real geometry	0.054531	0.014531	0.029398	0.032820
D_{aug}	0.052505	0.033093	0.039908	0.041835
S_{aug}	0.006018	0.003670	0.004477	0.004722

Table 42: **Matched prediction comparisons on screened additions.** NLL and differences are bits/target. Each row compares separately fitted arms under one training scheme; differences between rows are not effects of switching training schemes. Positive gains favor geometry.

Training / comparator	Comparator	Geometry	Gain	97.5% interval
Canonical / index	3.400790	3.360967	**0.039823**	[0.036257, 0.043532]
Raw / pseudo-coordinates	3.401170	3.368350	**0.032820**	[0.029216, 0.036302]
Augmented / index	3.393776	3.351940	**0.041835**	[0.036298, 0.048860]

I.1 Source vocabulary coverage

This post-hoc partition uses each study's actual source encoding: canonical for the leading-space study, raw for the pseudo-coordinate study. Known nonempty sources, sources containing input UNK (token 2), and empty sources contain 2,603,957, 56,715, and 1,915 targets, respectively. Output UNK (class 1099) and all target labels remain unchanged. Contributions use original population weights; conditional gaps divide by each layer's original mass, without re-equalizing records or blocks. Within each study, the three signed contributions add back to its primary contrast; no new intervals are assigned.

Table 43: Source-coverage decomposition, in bits/target. Canonical: leading-space canonical index minus geometry; Pseudo: pseudo minus real geometry. Contribution retains full-population weight; the mean and seed columns are conditional gaps. Seed suffixes 13–15 denote 20260913–20260915.

Study	Source	Mass (%)	Contribution	Mean gap	Seed 13	Seed 14	Seed 15
Canonical	Known	96.6554	0.038716	0.040056	0.055991	0.037623	0.026553
Canonical	UNK	3.1799	0.001152	0.036233	0.014734	0.051863	0.042104
Canonical	Empty	0.1647	−0.000045	−0.027244	−0.364166	0.122554	0.159880
Pseudo	Known	96.6554	0.033140	0.034287	0.056052	0.016470	0.030338
Pseudo	UNK	3.1799	−0.000415	−0.013040	0.013309	−0.044887	−0.007540
Pseudo	Empty	0.1647	0.000095	0.057610	−0.042267	0.023622	0.191474

Known nonempty sources carry 96.6554% of weight and favor geometry in every seed of both studies. The pseudo comparison's UNK layer is negative on average and remains included in its positive total.

J Equivalent encodings and prediction stability

Rewrite and exact canonicalization. In a nonempty-ink source's leading prefix of U+0020 and U+3000 spaces, one step exchanges five ideographic spaces for eleven ordinary spaces, or the reverse when available. Their MS PGothic 16px advances are 11 and 5px, so the exchange preserves 55px. Required spaces are consumed left to right, replacement is inserted at the first consumed position, and other characters retain their order. Target history, coordinates, and labels stay fixed; source positions are recomputed. Bound whole-row pixel checks verified every tested view under the specified font and BASIC rendering configuration: 101,762 comparisons in the earlier recoding/canonical-input study and all 29,630 two-step source rows in the augmentation study.

For prefix width W, the canonicalizer chooses nonnegative (a, b) minimizing $(a + b, a, b)$ lexicographically subject to $5a + 11b = W$, then writes b ideographic spaces, a ordinary spaces, and the unchanged suffix. Thus $C(R) = C(T)$ for this rewrite, giving identical deterministic predictions and JS$= 0$. This is a property of the input map. The earlier canonical-input experiment changed inputs to frozen raw-trained models; the canonical-training study in Appendix I instead normalizes fitting, development, and evaluation.

Augmented models on common support. One-step expansion covers 1,442,509 additions targets and 55.6580% of original weight. Two-step support contains 769,265 targets in 2,450 records and 555 of 611 blocks, carrying 28.7903% of weight. Table 44 fixes that latter support and its original hierarchical weights across all views. Base-2 JS is computed per target and seed from FP32 logits, using FP64 probabilities and 10^{-4} smoothing including fixed output-UNK probability, then averaged. It is not computed from a mixture of seed distributions.

Table 44: Augmented models on identical two-step support. NLL is bits/target, JS is bits, and accuracy/flips are percentages. Means average the three fixed seeds. Companion endpoints are descriptive; the two-step JS contrast is a primary quantity.

Model	View	NLL	Accuracy	JS vs raw	Flip vs raw
Index	Raw	3.309143	51.713	0	0
Index	One step	3.311653	51.663	0.006066	6.806
Index	Two steps	3.325298	51.523	0.012890	9.872
Geometry	Raw	3.272471	52.149	0	0
Geometry	One step	3.275044	52.169	0.004862	6.078
Geometry	Two steps	3.283344	52.052	0.008168	7.948

Geometry has lower JS, flips, and absolute NLL fluctuation in every paired seed at both measured intensities. Defining Γ_k as the index-minus-geometry difference in raw-to-k net NLL change, one-step values are -0.002271, 0.003100, and -0.001015 bits (mean -0.000062); two-step values are 0.004192, 0.010754, and 0.000901 (mean 0.005283). Small mean net changes can therefore coexist with substantial movement: one-step mean absolute NLL changes are 0.172958/0.153506 bits for index/geometry. Absolute NLL and accuracy favor geometry in all three views and seeds, but those comparisons do not identify why its distributions change less. Complete logits were not saved, so these outputs cannot resolve an entropy-based explanation.

The increase in JS from raw–one-step to raw–two-step is smaller for geometry in all three pairs: index/geometry increments are 0.007175/0.003039, 0.005899/0.003209, and 0.007399/0.003672 bits. Flip increments likewise are 3.0975/1.7826, 2.6805/1.7927, and 3.4200/2.0326 percentage points. These compare two measured intensities in one rewrite family; the increments are not JS or flip rates directly between the one- and two-step views.

Earlier frozen-model comparison. This follow-up used the six raw-trained epoch-10 checkpoints; its analysis was fixed after raw outcomes were known, before transformed-additions scoring. Both available directions share each target's weight equally, then normalize over eligible mass. Coverage is 1,442,509 targets, 2,833 records, 585 blocks, and 55.6580% of original weight. Expansion carries 99.97531% of allocated weight. Compression has only 386 targets in 15 records and 12 blocks, with descriptive $\Gamma = -0.020874$; all also permit expansion. The pooled directed contrast is $\Gamma = 0.027193$ bits, with 97.5% interval [0.025389, 0.029020] and seed range 0.017117–0.032656. Its family also contains canonical-input NLL gap 0.030930 [0.026868, 0.034840]. Those intervals use 10,000 paired draws of all 611 blocks (PCG64 seed 20260917), recomputing eligible denominators after fixed-seed averaging. They form a separate two-comparison family from the later training controls.

Canonicalization changes raw source sequences for 34.9592% of eligible weight; only 0.04938% has shorter sequences. Within the changed subset, the frozen models' canonical NLL gap remains 0.032576 bits (seed range 0.020321–0.051339). This describes canonical input to the old models, not the effect of canonical training.

J.1 Behavioral interpretation and archived detail

For the earlier frozen models, eligible-population JS is 0.016426/0.009979 bits, flip rate 11.257/8.561%, and mean absolute NLL change 0.285586/0.218386 bits (index/geometry). Among targets both arms originally classify correctly, recoding error rates are 2.5817/1.5961%; that common subset carries 48.71–48.87% of eligible weight across seeds. These post-hoc descriptions favor geometry in each seed, without controlling confidence or providing independent confirmation. About two-thirds of flips in both the old and augmented comparisons are between wrong labels, so flips are not additional errors.

The legacy reports listed in Appendix L.4 retain the full earlier tables and numerical audit trail. Cross-study differences between the old and augmented models do not isolate the effect of augmentation; the matched comparison is between arms within each study.

K Descriptive checks on the existing predictions

The earlier F-series checks below were designed after the reported results were known. They reuse saved predictions, keep the full target population and original weights, and introduce no new confidence intervals or primary tests. The machine-readable results and scope note are in `supplementary/descriptive-checks/`. The current all-run W analysis appears separately in Appendix A.5, with its own current-package entry point.

K.1 Sources actually changed by canonicalization

Changed status is determined from the source token sequence before and after the fixed canonicalizer, not from model performance. Recomputing this status on 83,479 source rows identifies 28,384 changed rows; target expansion gives 749,766 changed and 1,912,821 unchanged targets. Independent reductions agree. The models here are *trained on canonicalized sources*; these counts and weights do not describe the earlier raw-trained frozen canonical-input check.

Table 45: Canonical-training index-minus-geometry difference partitioned by source change. p_S is original estimand mass; contribution B_S preserves all denominators, and B_S/p_S is the conditional difference (bits/target).

Source tokens	Targets	p_S (%)	B_S	B_S/p_S
Changed	749,766	23.3428	0.008020	0.034359
Unchanged	1,912,821	76.6572	0.031803	0.041487
All	2,662,587	100	0.039823	0.039823

Changed-source conditional gains in the three fixed pairs are 0.049988, 0.026605, and 0.026484. Thus the overall gain is not confined to rows left unchanged by this preprocessing. The two strata are not randomized interventions; their contrast does not identify a causal interaction with canonicalization. Records and blocks may contribute targets to both strata.

K.2 Distribution across source blocks

Table 46 describes the seed-averaged within-block prediction differences on the same 611 blocks. Every study retains its original within-record and within-block target weights. Positive blocks are not independent new replications, and a positive three-seed mean does not imply that every seed or target improves.

The canonical comparison is positive in 89.53% of blocks after averaging fits; individual seeds are positive in 529, 507 and 476 blocks. Across the four studies, 460 blocks have positive seed-averaged differences in all four, and 540 in at least three; ten are negative in all four and three are zero. This measures spread within reused material, not independent corpus transfer or a ranking of training strategies. The complete signed block summaries remain in the machine-readable report.

Table 46: Descriptive breadth of positive index/pseudo-minus-geometry differences. The base row is frozen raw-trained F-index transfer; other rows are their separately fitted studies. Block medians are bits/target.

Study	Median	Positive	Negative	Zero	Positive in all seeds
Base index	0.033025	541	67	3	340
Canonical index	0.040691	547	61	3	370
Raw pseudo	0.033112	531	77	3	296
Augmented index	0.040673	559	49	3	—

K.3 Natural character–rendered position mismatch

This separate descriptive analysis uses the original raw-trained, frozen epoch-10 F-index/F-geom pairs on all screened additions, not the canonical-trained models or the recoding-eligible subset. Its rules were fixed after the overall transfer results were known and before inspecting the stratum effects. For target column c and source length L, define $\delta = |g_c - q_{\mathrm{MS}}|$ when $0 \le c < L$, using the common 16px width table and $+0.5$ anchor. This compares source and target positions at the same character ordinal; it is not a measured attention or alignment error.

The fixed strata are 0, $(0, 8]$, $(8, 16]$, and > 16px, plus empty sources and nonempty sources with $c \ge L$. All 2,662,587 targets and their original target/record/source-block weights remain in the partition. For three-seed mean NLL difference d_t, a stratum has original mass $m_S = \sum_{t \in S} \omega_t$ and contribution $k_S = \sum_{t \in S} \omega_t d_t$; the contributions sum to 0.032244 bits/target. The > 16px stratum carries 41.202% of this population's weight and contributes 0.017624 bits/target, or 54.658% of the gain.

Shared source-length/target-column cell weights give conditional gains of 0.013978, 0.029733, 0.026701 and 0.040980 bits/target in the four defined-offset strata, respectively. Source-length bins are 1–16, 17–32, 33–64, 65–128 and 129+; target-column bins are 0–7, 8–15, 16–31, 32–63 and 64+. A retained cell requires at least 20 targets, five records and three blocks in every stratum. The 19 retained cells cover all defined offsets (91.620% of original weight). Within-cell weights preserve the original hierarchy; pooled cell masses provide the shared reference. Thus the largest mean remains at > 16px, without a monotone four-bin pattern. This association does not identify the model's mechanism or remove all input-composition differences. Full partitions, per-seed values, rules and numerical replay are in `offset-analysis.zip`, under `archive/results/`, `protocol/analysis.json` and `evidence/verify_offset.py`.

L Inference scope, boundary results, and archived evidence

L.1 Shared inference scope and material reuse

The all-run experiment, its fixed-recipe repeats, representation controls, pseudo-metric, residual-dropout and larger normalized diagnostics, and the causal-capacity experiments share the previously examined ordinary additions: 2,951 records, 2,662,587 targets and 611 source-file blocks. Their design followed earlier findings; this is reused evaluation material, not blind data confirmation. The old neural evaluation has 632 records, 497,645 targets and 573 near-duplicate components. The W stress population has 512 components, 1,069 source rows and 6,493 targets, with component–row–target weights. Each population retains its own denominator. The separately acquired event material has 19 screened artworks, 24,748 targets and 18 visible-signature groups. Its inclusion rules, models and analysis were frozen before inference. It reuses the additional stage's three fitted pairs, adds no training repeats, and is reported separately from the ordinary additions (Appendix A.3).

Stress selection froze inputs against a finite ledger of prior uses, whole-record relations and exclusions, including P1, a sealed held-out pool excluded from all reported results. This controls recorded overlap rather than certifying author, source-file or local-template independence. Later source-mask and capacity comparisons reuse the stress material and three paired seed identities. They are separate comparisons, not additional independent samples or a complete factorial design. The new larger

normalized comparison is bidirectional (Appendix B.3); no larger all-run-normalized causal model was evaluated.

Each material interval averages the stage's three fitted pairs before paired block/component/group resampling, conditional on fits, inputs, groups and selected endpoints. It excludes retraining, pilot choice, endpoint selection and research adaptation. Descriptive seed-t intervals instead summarize the mean of three paired differences under a normal-difference assumption that cannot be validated with $n = 3$. Both interval types, including those crossing zero, are retained. Table 47 distinguishes original endpoint families from their roles in this paper.

The additional all-run stage uses three new seed identities at fixed epoch 8 without development reselection. It remains separate from the original three pairs used to select that endpoint: these are two training stages on the same material, not one prospectively fixed six-pair experiment.

The current W pseudo-metric and residual-dropout diagnostics (Appendix B) each specify one ordinary NLL contrast with an individual 95% material interval and descriptive fit-mean interval. Both arms are newly trained within each diagnostic, reusing the additional stage's initial states, seed identities and data orders at fixed epoch 8. These are two separate diagnostic families rather than additional independent-seed repeats or a joint-coverage claim.

Leading-space canonicalization guarantees identical predictions only for its specified equivalence classes; all-run normalization extends this guarantee to maximal U+0020/U+3000 source runs, not arbitrary visual equivalence. The earlier raw recoding population is 99.975% expansion by allocated weight, with a negative descriptive compression-stratum loss-increment contrast. The earlier F-pseudo study uses one fixed width permutation within a shared real-target metric and has a negative source-UNK stratum. Half-data seed signs are mixed. Larger causal d256 does not improve over d128 in mean ordinary NLL, and some seed-level right-side and proportional stress effects are unfavorable. These outcomes remain in their relevant tables. No branch identifies internal prefix-sum computation, validates attention alignments, or measures free-generation layout quality or second-font full-row transfer. Moving older evidence into an archive changes neither its inference family nor its status.

L.2 Original local counting confirmation

This earlier experiment concerns externally selected local source access, not the later complete-row representation comparison. A smoothed counting predictor uses two preceding target-row characters and one character from the previous row, selected by character column (C) or MS PGothic rendered coordinate (P). Eligible drawings have at least 15 rows, maximum character width at least 16, and no tabs; padded canvases and eligible nonspace targets define this experiment's population. Five-fold component-blocked selection chooses among 56 count-mixture weights before independent confirmation. The confirmation retains 661 records, 518,874 targets, and 603 component–height units from 600 near-duplicate components. Targets average within units, represented height strata within components, and components equally.

Let M, N, F, NF denote mean $\log_2$ scores for adjacent, query-permuted adjacent, non-adjacent, and query-permuted non-adjacent channels, respectively. Five paired replicas define each null/floor channel. With $G_w^{\mathrm{adj}} = (M_w - N_w) - (F_w - NF_w)$, the original confirmed difference is

$$\Delta^{\mathrm{adj}} = G_P^{\mathrm{adj}} - G_C^{\mathrm{adj}} = 0.122043 \quad [0.112915,\ 0.131171].$$

This interval belongs to the original eight-estimand simultaneous 95% construction using 2,000 paired component draws, not to eight independent replications. A later frozen-score replay gives aligned NLLs 4.155440 (C) and 4.025951 (P), a direct improvement of 0.129489 with a *post-hoc descriptive* 95% interval [0.121262, 0.138204]. These are different estimands. The replay retains the original targets and weights without refitting and does not acquire the original confirmation status.

The original target-column null can depend on target or later-character widths. A post-hoc fixed source-address cyclic-shift null retains an adjusted advantage of 0.123972 under the primary selection scheme, while leaving aligned/floor probabilities unchanged. This changes the null distribution and estimand; it does not identify a causal leakage contribution or certify a complete autoregressive interface. All eight original estimates and intervals, decision rules, 18 planted-label calibration conditions, absolute scores and alternative-null reports remain in the companion reports indexed below.

Table 47: Original inference families and present role. Nominal coverage is within each stated family; it is not joint coverage across this research sequence.

Study	Endpoint, interval and population
All-run W normalization	One original primary ordinary NLL contrast; 95% paired interval from 10,000 resamples of all 611 additions blocks. Main prediction evidence here.
Additional all-run repeats	One prespecified NLL contrast across three new pairs at fixed epoch 8; 95% material interval over the same 611 blocks, with a separate descriptive mean fit interval. Original pairs are excluded from this stage's estimate.
Fixed-recipe W representation controls	Width-only−Scan is the sole prespecified primary NLL contrast; Linear−Scan is secondary and descriptive. Each has a 95% material interval and descriptive mean fit interval. The existing three Scan fits and all 611 blocks are reused; no joint coverage is claimed.
Frozen-model event evaluation	One prespecified Index−Scan NLL contrast on 19 screened works in 18 visible-signature groups; 95% conditional paired group-bootstrap interval and descriptive mean fit interval. The existing additional-stage models are fixed; no new training repeats or pooled ordinary-material estimate.
Event representation controls	Post-hoc follow-up on the same 19 works and existing models: Width-only−Scan is primary within this diagnostic, Linear−Scan secondary. Separate 95% conditional material and descriptive mean fit intervals, with no joint confirmation claim.
Larger normalized regularized W	One primary Index−Scan NLL contrast with an individual 95% material interval and descriptive fit interval, using three new paired seeds and one common development-selected endpoint. Capacity and recipe change jointly; all 611 ordinary blocks are reused.
Causal d128 and d256	One original primary ordinary NLL contrast per capacity, each with a 95% paired interval over 611 blocks. Later 97.5% per-capacity intervals are a descriptive two-capacity Bonferroni sensitivity, not an original joint family. Main prediction evidence here.
Original and matched-index F	Each old-material comparison has its own 95% paired component interval; the matched-index extension reuses the F-geom trajectory.
Frozen F additions transfer	F-char/F-geom and F-index/F-geom each have a separately specified 95% paired source-block interval. The latter followed the old-index and original-transfer results.
Leading-space and pseudo training	Two NLL contrasts, each 97.5%, from 10,000 shared draws of all 611 blocks.
Augmented F training	Full-population NLL and eligible two-step JS are a separate two-contrast family, each 97.5%; all 611 blocks remain, including 56 with zero eligible mass.
Earlier frozen-model recoding	Directed loss-increment contrast and canonical-input NLL gap form a two-contrast family, each 97.5%; all blocks remain with sampled eligible-mass denominators recomputed.
Reduced training material	Three additions contrasts, each 98.3333%; old-domain and fixed-update comparisons are descriptive.
Original W masks and larger bidirectional W	Each original study designated K_2 primary, with a 95% paired interval over 512 stress components; ordinary NLL was secondary. Those designations and intervals are retained. These results provide supporting behavioral evidence here, without a joint mask/capacity test.
Larger causal stress outcomes	K_2, side-specific JS, accuracy, flips and entropy are secondary; T is a later proportional-scale description. Their intervals do not replace the primary ordinary-NLL comparisons.
Coverage and behavioral partitions	Known/UNK, offset, prediction-change and common-correct results are descriptive, with no new subgroup tests.
Original counting study	Eight-statistic simultaneous 95% confirmation; the later direct-loss interval is a separate post-hoc description.

Calibration applies to the counting instrument, not the neural models or universal natural-null error control.

L.3 Reduced training material

This boundary study selects 290 of 581 fit components by a fixed seed-dependent hash and retains their targets. F-index/F-geom are paired within each of the three original seeds. Full-fit vocabulary and geometry scales remain available; index moments are fitted on each subset. Thus the intervention reduces gradient-update material, not all preprocessing information. Each trajectory completes 168,000 updates. One six-model mean development loss selects the common half-data endpoint from 8,400/16,800/42,000/84,000/168,000 updates, choosing 16,800; full-data models retain their original 84,000-update endpoint. All additions targets and original hierarchical weights remain scored.

Table 48: Reduced-data boundary on additions (bits/target). The three primary intervals use 10,000 shared paired draws of the 611 source blocks, PCG64 seed 20260917, and Bonferroni correction, conditional on fits, subsets and endpoints.

Training material	F-index NLL	F-geom NLL
Full	3.396202	3.363959
Half	3.593299	3.585271
Primary contrast	**Estimate**	**98.3333% interval**
Full index − half geometry	-0.189069	$[-0.204694, -0.175684]$
Half index − half geometry	$+0.008028$	[0.004557, 0.011801]
Half gain − full gain	-0.024216	$[-0.029569, -0.018801]$

Half-data paired gains are $+0.001271$, -0.002939, and $+0.025752$; their signs are not uniform. At the separately specified 84,000-update reference, half-data index/geometry NLLs are 3.924016/3.897929: the larger descriptive gap of 0.026087 accompanies worse absolute losses in both arms. It does not replace the development-selected endpoint. These results do not establish a halving of training-material requirements or larger benefits with less data; two data amounts do not identify a scaling law. The complete old/additions, seed, and fixed-update values are retained at the legacy-packet locator `supplementary/reduced-data/`.

L.4 Earlier exploratory studies

Earlier local-access experiments, other-font explorations and tokenization checks are separate from the complete-row W comparisons. Their archived reports retain positive, mixed, negative and incomplete outcomes: NetHack results depended on predictor and endpoint; Osaka results did not establish second-font complete-row transfer; fixed-offset tokenizer failures did not establish model inability; Qwen S3 had no established functional gain; Gemma did not pass its complete-block gate; and T-syn did not pass its registered screening conjunction. The reserved P1 pool remains excluded from all reported results. These branches provide no additional W fits, populations or confirmation claims.

Their two report locators, `supplementary/previous-evidence-report.pdf` and `supplementary/evidence-archive.pdf`, refer to separately retained legacy assets, not files or download addresses in this package. The current appendices retain the methods, numerical comparisons and boundary results used in this paper; executable entry points are listed below.

L.5 Reproduction entry points

The package's top-level README maps each result to its executable entry, dependencies and commands. The entries below distinguish reconstruction of inputs, fresh model inference and replay of saved statistics. The delivery checks do not repeat training. Numerical prediction uses the frozen advance table. Pixel checks additionally require the matching font and renderer.

HukuTemp inputs and the original normalized W models. `allrun-reproduce/` supplies the six original-stage bidirectional $d64/2$ parameter dictionaries for Table 1, the model implementation, vocabulary, advance table and frozen calibration. The 8,455,044-byte exports bind all models to the selected common epoch 8 / step 67,200; optimizer states and natural prepared arrays are not included. `reconstruct_unicode.py` uses the exact HukuTemp v32.2 ZIP and the included `metadata/records.json` to recover all 2,951 Unicode records with their original file hash. The 3,609,744-byte manifest retains its original hash, representative members and block identities. Reconstruction preserves historical parsing, entity decoding and normalization, and recovers the fixed selection and grouping without rerunning their decision process. The upstream archive is external; `UPSTREAM_RECONSTRUCTION.md` gives its release post, full-download page, size, hash and commands.

`rebuild_inputs.py` constructs canonical sources, histories, queries, targets and weights from those records. All 18 arrays match their original SHA-256 hashes across 2,662,587 targets, without loading archived numerical inputs or weights. Checks cover all-run canonicalization, nonspace pen origins, positive-width accumulation and width lookup before UNK mapping; changing a target character or future suffix leaves its prediction inputs unchanged. A clean-directory archive-to-array reconstruction precedes the declared eighteen-model forward subset. `run_forward.py` checks input schemas, hashes and weights, then computes fresh NLL and accuracy without reading archived target scores. Missing or changed assets produce explicit diagnostics.

The original six-model CPU check uses 512 target IDs uniformly spaced in the original order, fixed without consulting predictions. All 3,072 argmax choices and correctness flags match CUDA; maximum NLL difference is 1.96061×10^{-5} bits against a predeclared 10^{-4}-bit tolerance. Batch inputs match the production builder tensor for tensor. This delivery check does not repeat the full 2,662,587-target forward run. `run_smoke.py` executes all six models on three self-authored examples, checking software execution rather than predictive performance.

Four-arm normalized controls. `representation-controls/` supplies twelve dictionaries: the three additional-stage Index/Scan pairs and three Width-only and three Linear models. It includes frozen Linear calibration, feature and input builders, portable inference and block-statistic replay; receipts distinguish reused references from new fits. Clean-directory feature checks cover constant-width rows, single characters, empty/NULL sources, PAD, UNK and long sources. At 512 fixed natural targets per model, all 6,144 choices match archived scores, with maximum NLL difference below 1.93×10^{-5} bits. This is a forward subset check, not repeated population inference or new-material confirmation. Both entries carry the same original record manifest and share the archive-to-text path above. Self-authored smoke inputs and saved block arrays are included; natural texts and arrays are reconstructed from the external archive.

Public-source event evaluation and controls. `kouhaku-reproduce/` supplies public forum/catalog locators, fixed artwork boundaries and hashes, extraction, input construction, frozen-model scoring and aggregate replay. It loads the six additional-stage Index/Scan dictionaries from `representation-controls/`. Clean online reconstruction matches all 19 texts and 19 input arrays exactly. An initial CPU check on 512 targets per model reproduces all 3,072 CUDA choices, with maximum NLL difference 1.9606×10^{-5} bits. A later clean public-download/input/full-forward chain reproduces all 148,488 target–model choices, with maximum difference 4.5920×10^{-5} bits. Separate receipts identify both software checks and their distinction from material confirmation. Protocols, grouping, full-score receipts and per-artwork summaries accompany this entry (Appendix A.3). Downloaded HTML, source texts and natural arrays are not redistributed; changed sources fail identity checks.

`kouhaku-controls/` scores the existing six Width-only/Linear models on those same event inputs. It includes the frozen branch and calibration, all-work statistics, completion/integrity receipts and four-arm replay. All six full forward evaluations preceded aggregation. A separate 512-target-per-model check matches all 3,072 choices with maximum NLL difference below 5.5×10^{-6} bits. Replay retains every original Index/Scan leave-one-group-out estimate and both alternative weighting summaries. These are follow-ups on the same event frame (Appendix A.4).

Larger normalized models and normalized diagnostics. `larger-normalized-reproduce/` supplies six selected FP32 models, configuration,

saved-block replay and a scorer using the all-run builder. Appendix B.3 documents common epoch-6 selection, independent full-score reaggregation and fixed 512-target checks; the delivery check does not repeat ordinary population inference. `normalized-diagnostics/` supplies twelve fixed-epoch models, features, calibrations and separate replay for the pseudo-metric and residual-dropout contrasts (Appendix B). Its `input_bridge/` verifies the complete source-bank identity before vocabulary mapping and constructs original-codepoint inputs from the shared records/arrays. Complete statistical replay and 512-target-per-model forward checks are distinct; commands also support fresh inference. Reused seed identities remain reused, not additional independent repeats.

Causal-model demonstration and numerical replay. `core-reproduce/run_demo.py` executes the selected $d = 256$ recipe-B/epoch-4 pair for seed 20260919 (3,924,555 parameters per arm) on two fixed self-authored queries. Strict loading, feature reconstruction, cached/reference inputs and logits, and the saved CPU reference all pass on the documented runtime. One rewrite is more disruptive for Scan and is retained. These leading-run fixtures do not implement bilateral internal-space stress or reproduce K_2, and provide no natural-population, pixel-equality or CPU/CUDA bitwise-equivalence result.

`supplementary/new-results/replay_statistics.py` uses NumPy to replay four ordinary NLL results (two all-run stages, two causal capacities) and two larger-causal secondary K_2 results from saved block/component arrays. It verifies packaged hashes and ordinary bootstrap replicas without model execution or target-score reaggregation. The accompanying README indexes all-run, causal-capacity and larger-bidirectional arrays, exact summaries, interval specifications, all candidate development records, per-seed values and receipts. Its `source_manifest.json` separates original source hashes from packaged hashes. Logical relative identifiers replace personal filesystem roots while preserving numeric values and embedded hashes; these provenance identifiers are not download addresses.

Legacy evidence and asset boundaries. The separately archived legacy packet's `scripts/main_results.py`, `replay-README.md` and `supplementary/entrypoints.json` index earlier F, leading-space, causal-W and bidirectional-W score replay, with the corresponding code, advance tables and block statistics. Its width-scan, W-additions, bidirectional and canonical-forward entries distinguish saved-score replay from selected-checkpoint inference. The historical canonical forward check covered all 2,662,587 additions targets for six models at step 84,000, with unchanged weights and score parity; it adds no fit, population or inference family. The legacy `aa-real-model-diagnostic.zip` is a single-seed, self-authored-text diagnostic. These locations are in the legacy packet, not the current ZIP.

The current package includes the selected parameter exports described above, but not full optimizer/training checkpoints, natural source texts, prepared source-bank arrays or proprietary fonts. HukuTemp reconstruction requires the specified archive and included manifest; event reconstruction uses public pages. Saved-score replay verifies arithmetic and identities; population inference additionally needs model code, selected weights and exact inputs. Existing execution and reaggregation receipts retain their stated scope rather than implying a new from-scratch replication.