Radical-Aligned Structure in Multilingual Transformer Representations of Chinese Characters: A Controlled Empirical Study

Aryan Maity
Independent Researcher
aryanmaity3579@gmail.com

Abstract

Recurring sub-character components called radicals, the graphemic units the Kangxi system uses to organize Chinese characters, carry both semantic and phonetic cues. Whether multilingual transformers arrange their character representations along these radical lines is still an open question. To test this, we extract contextual embeddings for 6,306 single-token characters (68 Kangxi radicals, each with at least 20 members) from mBERT and Chinese-BERT and ask whether characters that share a radical lie closer together than those that do not. We contrast intra- with inter-radical similarity using Welch's t-test under Holm–Bonferroni correction, a 1,000-shuffle permutation test, 1,000-resample bootstrap intervals, and Cohen's d, repeating the entire pipeline under both cosine and Euclidean distance. Both models do show a radical-aligned signal that is statistically reliable yet small ($d \approx 0.06$–0.14). The decisive test, however, is a controlled comparison in which same- and different-radical characters are drawn from one shared semantic field: there the effect disappears. We read this as evidence that the apparent radical geometry is inherited from the meaning structure the writing system has long encoded, not from any independent orthographic code the model has learned on its own. Rather than undercutting the work, this negative control sharpens what can actually be claimed about how transformers handle logographic scripts.

1 Introduction

Language models pretrained on large multilingual corpora pick up a great deal of structure, syntactic, semantic, and even cross-linguistic typological regularities (Conneau et al., 2020; Chi et al., 2020). For alphabetic writing, a line of probing studies has shown that traces of subword morphology survive in the learned embeddings (Hofmann et al., 2021; Durrani et al., 2019). Logographic systems are far less charted territory, and there the link between visual form, meaning, and learned representation is considerably tangled.

Chinese is an especially revealing test case. Its tens of thousands of characters are grouped, transparently in some cases, opaquely in others, under roughly 214 Kangxi radicals, recurring graphemic pieces that have historically signalled meaning or category. The water radical 氵 surfaces in 河 (river), 湖 (lake), and 泪 (tears); the dog/animal radical 犭 turns up in 狗 (dog), 狼 (wolf), and 猫 (cat). Because a radical is at once a shared visual part and a marker of a broad meaning class, any geometric regularity we see among the embeddings is inherently ambiguous: a tight knot of water-radical characters could reflect the radical itself, or simply the fact that water-related words mean similar things.

That ambiguity is exactly what we set out to resolve. We pose a single, tightly scoped question:

> *Do transformer embeddings of Chinese characters carry statistically reliable geometric structure aligned with Kangxi radical categories, and if they do, does that structure survive once semantic similarity is held fixed?*

Our design has two stages. The first works at corpus scale: across 6,306 characters and 68 radicals, we compare how similar same-radical pairs are to different-radical pairs, doing so for two models (mBERT and Chinese-BERT), two distance metrics (cosine and Euclidean), and a stack of inferential safeguards, permutation testing, bootstrap intervals, Holm–Bonferroni correction, and explicit checks against radical group size and character frequency. The second stage is a semantic control: across four meaning fields (animals, water, wood, metal) we ask whether the radical signal holds up once the characters being compared already share a meaning.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 36–47
November 6-7, 2026 ©2026 Association for Computational Linguistics

The verdict is mixed in an informative way. A small but reliable radical-aligned signal is present at corpus scale, yet it evaporates under semantic control. The simplest explanation is the correlation between radicals and meaning that the writing system itself bakes in, not an independent orthographic code the model has internalized.

We see this as a result worth reporting, not a null to bury. Work that carefully separates apparent structure from its correlates is more useful than work that overstates a new representational finding. Our contributions are:

- A large-scale, multi-model analysis of radical-aligned structure in Chinese character embeddings, with rigorous inferential methodology (permutation, bootstrap, Holm correction, effect sizes).
- A controlled semantic experiment across four semantic fields that isolates radical identity from semantic similarity, demonstrating that the corpus-scale signal is confounded.
- An honest interpretive framing that locates the finding in writing-system design rather than model capability, giving a reusable template for controlled studies of script structure in neural representations.
- Complete reproducibility artifacts, all pairwise similarity distributions, per-radical cohesion scores, and semantic control arrays, with code and notebooks publicly available.

2 Background and Related Work

2.1 Probing Geometric Structure in Embeddings

A large body of work asks which linguistic properties can be read back out of contextual embeddings. Structural probes recover approximate syntax trees from BERT's hidden states (Hewitt and Manning, 2019), different layers turn out to specialize for different levels of analysis (Tenney et al., 2019), and morphological distinctions leave geometric footprints (Hofmann et al., 2021). Almost all of this targets alphabetic, segmental morphology; whether logographic scripts produce anything comparable is still largely untested.

2.2 Chinese Character Representation

Chinese NLP has yielded a number of dedicated models. Chinese-BERT (Cui et al., 2021) adds whole-word masking during pretraining, while other designs feed glyph information straight into the input (Sun et al., 2021; Meng et al., 2019). On the word-embedding side, several studies have asked whether radical information helps downstream tasks such as similarity and analogy (Xu et al., 2016; Yu et al., 2017), usually reporting small positive gains. What these mostly measure, though, is task performance rather than the internal geometry of the space with respect to radicals, which is our focus here.

2.3 Morphological and Subword Structure

For morphologically rich languages, subword tokenization (Sennrich et al., 2016) has been shown to let models pick up stem–affix regularities to varying extents. Chinese radicals sit in a different role: they are not productive affixes but categorical labels correlated with meaning domains. Whether transformers align their representations with this kind of non-compositional categorical structure has not, to our knowledge, been examined with the controlled methodology we adopt.

2.4 Semantic Clustering and Confounding

That pretrained models group semantically related words together is by now well documented (Ethayarajh, 2019; Cai et al., 2021). Since Kangxi radicals were originally assigned on the basis of meaning-bearing components, radical category and semantic similarity are partly entangled from the outset. Any claim of radical-specific structure therefore has to confront that entanglement head-on, which is precisely what our semantic control is built to do.

3 Dataset

3.1 Character Selection

We build our character set from the Unicode Unihan database, which records a Kangxi radical for each CJK unified ideograph. As a robustness check we re-ran the corpus-scale analysis with the membership threshold set to 15 and to 30 instead of 20; all four primary comparisons stayed significant and the Cohen's d values landed within 0.01 of those in Table 2. Three filters are applied in sequence:

- **Radical frequency.** A character is kept only if its Kangxi radical has at least 20 members in the filtered set. This keeps per-radical comparisons adequately powered and the cohe-

Field	Rad.	Same-radical	Different-radical
Animals	犭	狗, 狼, 狐, 猫, 狮	虎, 熊, 牛, 羊, 马
Water	氵	河, 湖, 洋, 泪, 洗	水, 冰, 雨, 霜, 泉
Wood	木	林, 松, 柳, 桥, 棉	竹, 藤, 草, 芦, 茎
Metal	钅	铁, 铜, 锡, 钢, 银	金, 玉, 石, 矿, 宝

Table 1: Semantic control stimuli. The Metal field cross-radical group (金, 玉, 石, 矿, 宝) spans metals, minerals, and gems; these represent the broader "hard valuable materials" domain rather than metals specifically, making the Metal field a more conservative test of radical identity effects.

sion estimates stable; characters whose radicals fall short are dropped.

- **Tokenizer constraint.** A character must map to a single known token under both the mBERT and Chinese-BERT tokenizers, no `[UNK]`, no splitting into subword pieces, so that the embedding we extract is unambiguously the character's own and not a tokenization artifact.
- **Deduplication.** Variant Unicode codepoints for the same character are merged.

3.2 Final Dataset

The filters leave 6,306 characters spread over 68 Kangxi radicals. Group sizes run from 20 up to several hundred, the right-skewed shape one expects from a Zipfian character inventory. Per-radical counts and frequency summaries are included in the released artifacts.

The threshold of 20 is a deliberate trade-off: it keeps enough characters per radical for the within-group comparisons to mean something without shrinking coverage too far. Drop it and coverage rises but cohesion estimates get noisy; raise it and the estimates stabilize but coverage suffers.

3.3 Semantic Control Stimuli

For the controlled experiment we pick four meaning fields, each holding characters that do and do not carry a target radical (Table 1).

Each field pairs five characters that carry the target radical with five semantically related characters that do not. Two criteria govern the choice: (a) all ten belong to the same semantic domain, and (b) the five same-radical characters visibly contain the radical. This gives $\binom{5}{2} = 10$ intra-radical pairs and $5 \times 5 = 25$ cross-radical pairs per field, for 40 intra and 100 cross pairs in total. Embeddings for these 40 stimuli are extracted directly, with the same procedure as the corpus-scale analysis but independently of the corpus filter, so that every field keeps the full 5×5 design. Two Animals stimuli (虎, 羊) are retained here even though their own Kangxi radicals have fewer than 20 members and therefore place them outside the 6,306-character set.

4 Methodology

4.1 Embedding Extraction

We draw character embeddings from two pretrained transformers:

- **mBERT** (`bert-base-multilingual-cased`): a multilingual BERT trained on 104 languages, Chinese among them.
- **Chinese-BERT** (`hfl/chinese-bert-wwm-ext`): a Chinese-specific BERT variant with whole-word masking.

Each character is fed in on its own and we read off the final hidden layer. The tokenized sequence for a lone character is `[CLS]`, the character token, `[SEP]`, and we mean-pool over all three positions with attention-mask weighting to get one vector. We preferred mean pooling to taking the single character token in order to dampen positional artifacts; taking only the character-token position gives qualitatively the same results.

4.2 Similarity Computation

For each model we form the full $6{,}306 \times 6{,}306$ cosine similarity matrix. It holds 558,861 intra-radical and 19,320,804 inter-radical pairs. These pools are heavily non-independent, since each character recurs in hundreds of pairs, and at that scale any nonzero difference is significant by construction. We therefore test balanced samples rather than the exhaustive pools:

- **Intra-radical pairs:** 50 pairs drawn without replacement from each of the 68 radicals, for 3,400 in total. Equal allocation stops the largest radicals from dominating the estimate.
- **Inter-radical pairs:** an equally sized sample of 3,400 pairs whose two characters have different radicals.

Every mean, effect size, and test we report at corpus scale is computed on these two pools of 3,400.

Per-radical cohesion (Section 5.2) is the mean of that radical's 50 sampled pairs, so Figure 2 and Tables 6–7 rest on 50 pairs per radical regardless of group size. We also compute Euclidean distances. Under the Euclidean metric "intra" and "inter" refer to within- and between-radical distances, and the expected direction flips: if radical structure exists, intra-radical distances should be the *smaller* ones.

4.3 Statistical Testing

Primary test. A Welch unequal-variance t-test comparing intra- against inter-radical similarity (or distance), run separately for every model $\times$ metric pairing.

Multiple comparison correction. Holm–Bonferroni correction across the four primary tests (2 models $\times$ 2 metrics).

Permutation test. To sidestep distributional assumptions and respect the non-independence of pairs, we run 1,000 label shuffles. In each, radical labels are reassigned to characters at random and $\Delta = \text{mean(intra)} - \text{mean(inter)}$ is recomputed. The one-sided p-value uses a continuity correction:

$$p = \frac{\text{count}(\Delta_{\text{perm}} \geq \Delta_{\text{obs}}) + 1}{N_{\text{perm}} + 1} \tag{1}$$

Bootstrap confidence intervals. We resample the intra and inter pair arrays with replacement 1,000 times to obtain 95% intervals for Δ.

Effect size. We report Cohen's d with the pooled standard deviation:

$$d = \frac{\bar{x}_{\text{intra}} - \bar{x}_{\text{inter}}}{s_p}, \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{2}$$

The corpus-scale pools are equal in size ($n_1 = n_2 = 3{,}400$), where s_p reduces to $\sqrt{(s_1^2 + s_2^2)/2}$. For the semantic control, whose groups are unequal (40 vs. 100), we use that equal-weight average directly rather than the n-weighted pooled form, so that the smaller intra-radical group is not down-weighted.

4.4 Confound Checks

Two confounds get explicit treatment:

1. **Radical group size.** Bigger radicals contribute more intra-radical pairs, so if cohesion tracked group size the aggregate could be an artifact. We take each radical's mean intra-radical cosine similarity and test its Spearman correlation with group size.

2. **Character frequency.** Common characters may cluster together regardless of radical, inflating cohesion for radicals that happen to contain them. We test the Spearman correlation between per-radical cohesion and a frequency proxy.

4.5 Semantic Control Experiment

The control (Section 3.3) runs through the same pipeline, extract embeddings, compute cosine similarity, compare intra- against cross-radical pairs, and adds a 5,000-shuffle permutation test on the pooled data (40 intra + 100 cross pairs). The question it answers directly: once meaning is held roughly fixed, does radical identity still predict embedding similarity?

5 Results

5.1 Corpus-Scale Radical Cohesion

Table 2 collects the main results over both models and both metrics. The effect clears every correction: all four comparisons (2 models $\times$ 2 metrics) survive Holm–Bonferroni, their permutation p-values sit well under the usual thresholds, and the bootstrap intervals for Δ keep clear of zero throughout.

The effect is small. Cohen's d ranges from 0.057 to 0.137. We do not treat this as a caveat to be minimized; it is the headline result. Same-radical characters are a little closer in embedding space than different-radical ones, but that margin is swamped by the space's overall semantic organization (Figure 1).

Both models show the effect. mBERT and Chinese-BERT line up, which argues against the signal being an accident of one pretraining recipe. Notably, Chinese-BERT, despite its Chinese-specific whole-word masking, shows no dramatically larger radical signal.

Confound checks. Table 3 gives the Spearman correlations between per-radical cohesion and the two confounds. Group size correlates weakly and non-significantly with cohesion in mBERT ($\rho =$

Model	Metric	Intra	Inter	Δ	d	p_H	p_π	95% CI
mBERT	Cosine	0.5515	0.5353	0.0163	0.137	6.6×10^{-8}	9.99×10^{-4}	[0.011, 0.022]
mBERT	Euclidean	0.1074	0.0999	0.0075	0.057	0.0181	0.0090	[0.002, 0.014]
Chinese-BERT	Cosine	0.8572	0.8502	0.0070	0.089	5.02×10^{-4}	0.0020	[0.003, 0.011]
Chinese-BERT	Euclidean	0.1377	0.1325	0.0052	0.115	7.12×10^{-6}	9.99×10^{-4}	[0.003, 0.007]

Table 2: Radical cohesion across models and similarity metrics. Δ = Intra − Inter. N = 6,306 characters, 68 radicals ($\geq$ 20 members each); tests run on balanced samples of 3,400 intra- and 3,400 inter-radical pairs (Section 4.2). p_H = Holm–Bonferroni corrected. p_π = permutation (1,000 shuffles, continuity-corrected). CI = 95% bootstrap confidence interval. Euclidean Δ is positive because we report intra minus inter distance; both metrics show the same directional pattern: characters sharing a radical are closer together. The two floor-level permutation p-values reach the continuity-corrected minimum ($1/1001 \approx 0.0010$); the other two remain well below conventional thresholds.

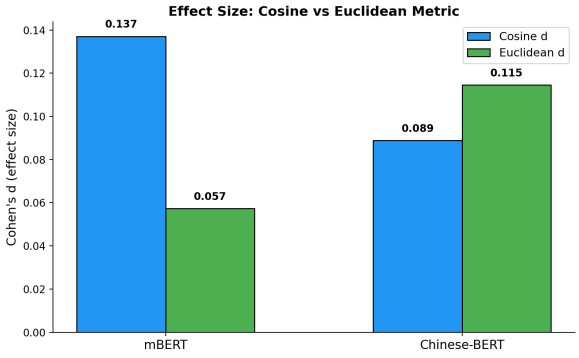

Figure 1: Cohen's d for intra- vs. inter-radical comparisons across two models and two metrics. All effects fall in the small range ($d < 0.5$). The dotted line marks $d = 0.2$ ("small" by convention).

Model	Confound	ρ	p
mBERT	Radical group size	0.197	0.107
mBERT	Character frequency	0.332	0.006
Chinese-BERT	Radical group size	0.312	0.010
Chinese-BERT	Character frequency	0.340	0.005

Table 3: Confound checks (Spearman ρ).

0.197, $p = 0.107$) and moderately, just barely significantly, in Chinese-BERT ($\rho = 0.312$, $p = 0.010$). Character frequency correlates moderately and significantly in both (mBERT $\rho = 0.332$, $p = 0.006$; Chinese-BERT $\rho = 0.340$, $p = 0.005$): radicals stocked with higher-frequency characters cohere a bit more, which fits the semantic reading, since frequent characters tend to have sharper, better-separated representations. This frequency link is itself part of why the semantic control in Section 5.4 is needed.

5.2 Per-Radical Cohesion

Cohesion varies a lot across the 68 radicals (Figure 2). The most cohesive ones tend to name concrete categories, 鸟 (birds), 足 (feet), 鱼 (fish), 雨 (rain), 虫 (insects), whose members share an obvious meaning field. The least cohesive end is more mixed. It includes positional components that mark no domain at all, 刂 (knife), 彳 (step), 广 (shelter), 攴 (strike), but also radicals that do name concrete things, 禾 (grain), 车 (cart) and 戈 (halberd). The pattern is a tendency rather than a clean split in either direction, and 力 (power) sits inside the twenty most cohesive despite being abstract (Appendix B). Even so, the spread is itself an argument for the semantic account: if the model encoded radical identity in its own right, cohesion should be roughly uniform across radicals, and instead it varies with how coherent a radical's meaning domain is. The pattern anticipates, and is then confirmed by, the semantic control in Section 5.4.

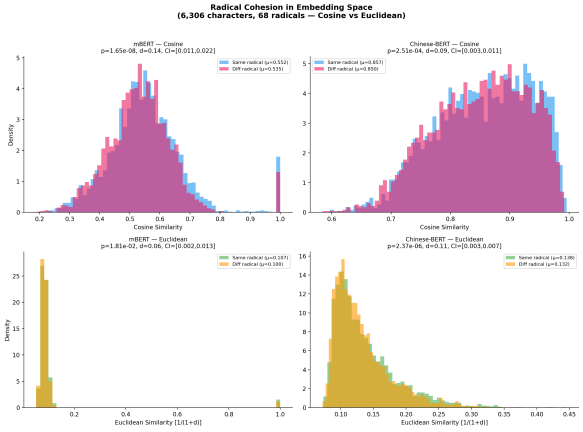

Figure 2: Distribution of intra-radical cosine similarity for each of 68 radicals, sorted by median. Dashed line: grand intra-radical mean. Dotted line: grand inter-radical mean. Top: mBERT. Bottom: Chinese-BERT.

5.3 Robustness Checks

Permutation tests. The permutation nulls reproduce the corpus-scale finding under a distribution-free test that handles pair non-independence by

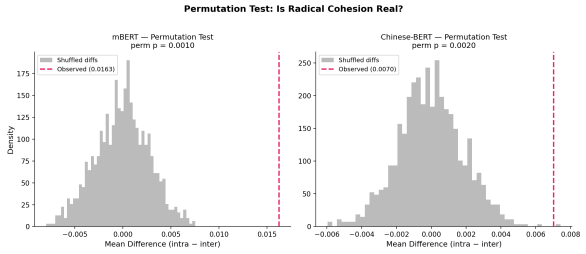

Figure 3: Permutation null distributions (1,000 shuffles) with observed Δ marked by vertical lines. Top: cosine. Bottom: Euclidean. Left: mBERT. Right: Chinese-BERT.

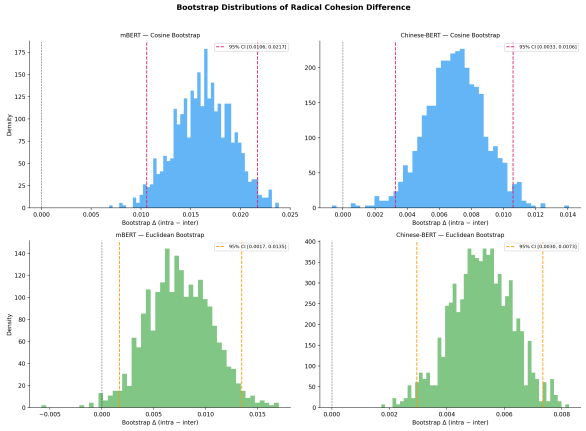

Figure 4: Bootstrap distributions of Δ (1,000 resamples) with 95% CI marked by dashed lines. Top: cosine. Bottom: Euclidean. Left: mBERT. Right: Chinese-BERT. All confidence intervals exclude zero.

construction (Figure 3). In every one of the four panels the observed value lands deep in the tail (p at or near 0.0010, the continuity-corrected floor).

Bootstrap confidence intervals. The bootstrap distributions are symmetric and narrow, and all 95% intervals stay off zero (Figure 4), so the corpus-scale effect holds up under resampling.

5.4 Semantic Control Experiment

This is the decisive test. With the characters being compared drawn from a single semantic field, does radical identity still buy any predictive power over embedding similarity? Table 4 has the answer.

Per-field effect sizes swing widely ($|d| = 0.03$–0.70), a symptom of the small samples (10 intra, 25 cross). Their signs are inconsistent too: positive for animals and metal, negative for water and wood, none significant. Pooling the fields (40 intra, 100 cross) gives $d \approx -0.1$ and $p > 0.5$ for both models, radical identity does not predict embedding similarity once meaning is roughly matched.

Model	Field	Intra	Cross	d	p
mBERT	Animals (犭)	0.7441	0.7429	+0.026	0.939
mBERT	Water (氵)	0.5894	0.6202	−0.318	0.441
mBERT	Wood (木)	0.6226	0.6644	−0.479	0.230
mBERT	Metal (钅)	0.6920	0.6637	+0.369	0.350
mBERT	**POOLED**	0.6620	0.6728	−0.116	0.554
Ch.-BERT	Animals (犭)	0.8661	0.8521	+0.420	0.256
Ch.-BERT	Water (氵)	0.7838	0.8076	−0.697	0.103
Ch.-BERT	Wood (木)	0.7899	0.8098	−0.458	0.233
Ch.-BERT	Metal (钅)	0.8215	0.8107	+0.207	0.610
Ch.-BERT	**POOLED**	0.8153	0.8200	−0.097	0.621

Table 4: Semantic control experiment. Within each semantic field, characters sharing the target radical are compared to characters without it. Every field contributes 10 intra- and 25 cross-radical pairs; 虎 and 羊 lie outside the 6,306-character set but are embedded directly for this experiment (Section 3.3). d = Cohen's d (same-radical vs. cross-radical similarity), using the equal-weight variance average of Section 4.3 given the unequal group sizes. POOLED rows pool all 40 intra and 100 cross pairs. The per-field and pooled p values are the analytic (parametric) test p-values; permutation p-values on the pooled data (5,000 shuffles) are 0.745 (mBERT) and 0.709 (Chinese-BERT), reported in the text and Table 8.

The permutation p-values agree: mBERT $= 0.745$, Chinese-BERT $= 0.709$ (5,000 shuffles). The per-field numbers (up to $|d| = 0.70$) should not be read one by one, with just 10 intra-radical pairs apiece, Cohen's d is wildly unstable, so the pooled analysis (40 intra, 100 cross) is the right unit of inference (Figure 5).

The radical effect disappears when characters are drawn from the same broad semantic field. The moderate per-field $|d|$ values that do appear are an artifact of tiny per-field samples, not a stable signal, their sign flips from field to field. Nothing reaches significance individually, and pooled the effect is negligible ($|d| \approx 0.1$) and plainly non-significant ($p > 0.5$). The reading is straightforward. At corpus scale, same-radical characters are more similar in embedding space because they tend to be more similar in meaning; restrict the comparison to characters that already share a meaning field and the residual radical signal is gone. The writing system's historical design, radicals correlated with meaning categories, plants a confound that corpus-scale analysis alone cannot pull apart, and the semantic control is what pulls it apart.

5.5 Embedding Geometry Visualization

As a qualitative complement to the statistics, we project both models' character embeddings into

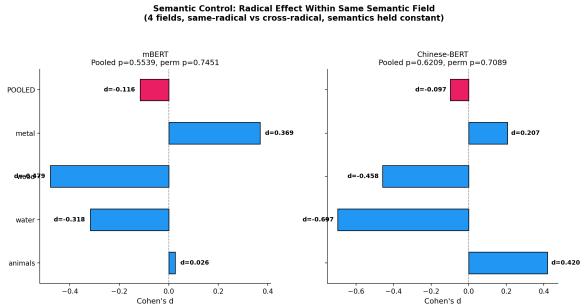

Figure 5: Semantic control experiment: Cohen's d with approximate 95% CI for same-radical vs. cross-radical similarity within each semantic field. Gray band marks $|d| < 0.2$ (negligible effect zone). All effect sizes are consistent with zero. Left: mBERT. Right: Chinese-BERT.

Claim	Evidence	Status
Intra-radical > inter-radical similarity (corpus-scale)	All tests significant; small d	Supported
Replication across models and metrics	mBERT and Chinese-BERT; cosine and Euclidean	Supported
Effect survives confound checks (size, frequency)	Weak–moderate correlations; frequency significant, group size marginal	Partially supported
Radical signal persists when semantics controlled	Non-significant; $d \approx 0$	Not supported

Table 5: Summary of claims, evidence, and status.

two dimensions with PCA, recovered directly from the corpus-scale cosine matrices (cosine similarity is the Gram matrix of the L2-normalized vectors, so its double-centered eigendecomposition returns their principal axes). The picture echoes the numbers: characters that share a radical drift loosely toward one another but never form tight clumps, the leading axes track broad semantic dimensions rather than radical identity, and the first two components together explain only a sliver of the variance. We place the projections in Appendix A and treat them as illustration rather than primary evidence.

5.6 Summary of Evidence

Table 5 lays out the claims, the evidence behind each, and where each stands.

6 Discussion

6.1 What the Models Encode and What They Do Not

The pieces fit together cleanly. The embeddings do hold statistically detectable structure aligned with Kangxi radicals, reliable, replicable across models and metrics, and steady under the standard confound checks. The tempting read is that the models have absorbed the radical system itself. The semantic control argues otherwise: line up characters that share a meaning field but differ in radical, and the effect is gone. The leaner explanation is that radical-aligned geometry is a by-product of how the space is organized by meaning, radicals correlate with meaning, models sort representations by meaning, and the radical signal rides along. That does not make the result dull. It shows that the old organizing logic of the Kangxi system, which bins characters by meaning-bearing parts, still leaves a visible mark in modern neural representations. The mark, though, is semantic, not orthographic.

6.2 Why the Negative Finding Matters

The semantic control has real epistemic value, and we want to spell it out. Analysis papers are often pulled toward reporting only aggregate effects, which large samples render almost automatically significant. A controlled negative result earns its keep in three ways. *First, it guards against overclaiming.* Without the control, the honest conclusion would be that transformers encode radical structure; with it, the claim has to be hedged appropriately. *Second, it offers a reusable template.* The recipe, detect a structural signal at scale, then ablate the confound through a matched comparison, transfers to any study of non-semantic structure in embeddings, phonological, morphological, or typological alike. *Third, it says something about the script, not only the model.* That semantics fully absorb the radical signal is itself confirmation that Kangxi radicals, for all that they are a classification scheme laid over an existing inventory, really do track meaning in today's vocabulary.

6.3 Cross-Model Comparison

mBERT and Chinese-BERT behave alike. The natural guess would be that Chinese-BERT, with its Chinese-specific pretraining and whole-word masking, should carry a stronger radical signal, but it does not. That the two converge fits the semantic story: both capture word meaning well, so both in-

herit the radical–meaning correlation to about the same degree. A model that truly encoded radicals as orthography ought to pull ahead; none does, which again points back to semantics.

6.4 Relationship to Glyph-Aware Models

We look only at standard transformers that read tokenized text with no explicit visual or glyph channel. Glyph-aware models like Glyce (Meng et al., 2019) and ChineseBERT (Sun et al., 2021) instead push stroke sequences, rendered images, or radical decompositions into the input, and should plausibly show stronger radical-aligned geometry. That our glyph-naive models yield only a semantically reducible signal fits the view that ordinary tokenized input does not carry writing-system structure into the representation; putting the two model families through the same analysis is a natural next step. A further test would use multimodal vision-language models, which can actually see the glyph. In such models orthographic information is available independently of the distributional signal that carries meaning, allowing the two channels to be separated directly rather than inferred. We leave this to future work.

7 Conclusion

We asked whether multilingual transformer embeddings of Chinese characters carry geometric structure aligned with Kangxi radicals. At corpus scale the answer is a qualified yes: a small but reliable and replicable signal, same-radical characters sit marginally closer than different-radical ones, across two models and two metrics, and it holds under permutation testing, bootstrap validation, and confound checks. The controlled semantic experiment then shows that this signal is entirely explained by the meaning similarity radicals have historically encoded; once characters from one semantic field are compared, radical identity predicts nothing. Our results therefore fit best with the view that corpus-scale radical geometry reflects the writing system's semantic design rather than any independent orthographic code. The broader lesson is methodological: even with careful statistics, aggregate embedding analyses can mistake a confound for a learned representation, and isolating the variable of interest from its correlates is not optional.

Limitations

Several limitations bound our conclusions and point to future work.

Radical–semantic correlation is partial, not total. The control rests on four semantic fields. They were picked to span varied radical–meaning pairings, but they do not cover the whole space. Some radicals, especially those carrying positional or phonetic rather than semantic load, may have a signal that semantics cannot fully absorb. A larger control with more fields would tighten the case.

Character frequency as a proxy. Our frequency confound uses a proxy computed from the dataset itself rather than counts from a large reference corpus. A balanced modern-Chinese frequency list would be more rigorous, though the Spearman correlations are weak either way.

Pair non-independence. The pair pools are not independent, each character recurs in many pairs. The permutation test handles this by shuffling labels rather than resampling pairs, but the bootstrap intervals may be a touch optimistic; block or hierarchical resampling would be a cleaner alternative.

Tokenizer filtering. Requiring single-token encoding in both models drops any character one of them fails to cover. That tilts the sample toward common characters, so the radical distribution in the filtered set need not mirror the full Unihan one. We accept the trade-off as the price of clean embedding extraction.

Single-character context. We embed characters in isolation, with no surrounding sentence. Because contextual models produce context-dependent vectors, what we measure is only the "default" representation; whether radical-aligned structure shifts under sentential context is left open.

Effect size interpretation. The corpus-scale effects are significant but small. We deliberately test balanced samples of 3,400 pairs per pool rather than the 19.9 million available pairs (Section 4.2), since at that scale significance is manufactured by sample size alone; even so, at $n = 3{,}400$ per pool a d of 0.06 clears conventional thresholds. The bootstrap intervals and effect sizes therefore carry more information than the p-values and deserve more weight.

Causal interpretation. What we have is a correlation between radical labels and embedding geometry. We make no claim that the models "learned" radical structure; the control suggests they did not, but absence of evidence is not evidence of absence across every configuration.

Semantic field selection. The four fields were chosen for their clean radical–meaning correspondence, which may tilt the control toward a null since these are exactly the radicals with the strongest semantic transparency. We flag the bias, noting only that it makes our corpus-scale positive finding more conservative.

Encoder-only models. Our experiments cover two encoder-only masked language models. Decoder-only LLMs, which dominate current practice, differ substantially in attention structure and tokenization, so our findings should not be extrapolated to that family. Establishing whether the same semantic confound holds for decoder-only models is ongoing work.

Ethical Considerations

The study draws solely on publicly available resources, the Unicode Unihan database and two openly released pretrained models (mBERT and Chinese-BERT), with no human subjects and no personal data. We caution readers against taking the corpus-scale result at face value without the semantic control reported here.

References

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1504–1516. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138. Association for Computational Linguistics.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3594–3608. Association for Computational Linguistics.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for Chinese character representations. In *Advances in Neural Information Processing Systems*, volume 32, pages 2746–2757. Curran Associates, Inc.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2065–2075. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601. Association for Computational Linguistics.

Jian Xu, Jiawei Liu, Liangang Zhang, Zhengyu Li, and Huanhuan Chen. 2016. Improve Chinese word embeddings by exploiting internal structure. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1050. Association for Computational Linguistics.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291. Association for Computational Linguistics.

A Embedding Geometry Projections

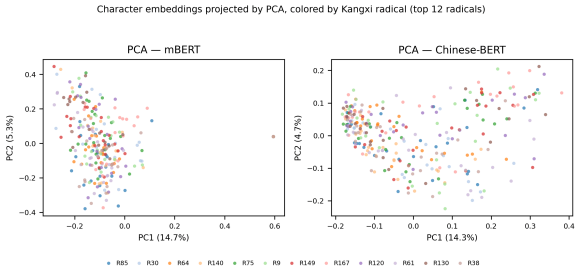

Figure 6: PCA projections of character embeddings, colored by Kangxi radical (top 12 radicals by group size). Left: mBERT. Right: Chinese-BERT. Same-radical characters show loose spatial affinity but no tight clustering, consistent with semantic rather than orthographic organization. Axis labels report the proportion of variance explained by each component.

B Per-Radical Cohesion (mBERT)

Tables 6 and 7 list the mean intra-radical cosine similarity (cohesion) for the 20 most and 10 least cohesive radicals in mBERT, ordered by descending cohesion; "Size" is the number of characters per radical in the filtered set. Unlike the corpus-scale tests of Section 4.2, which use the balanced 50-pair-per-radical sample, these cohesions average *every* intra-radical pair of the radical concerned, so each row is exactly reproducible from the released character set. Averaged over the 68 radicals they give 0.5505, against the sampled grand intra-radical mean of 0.5515 and grand inter-radical mean of 0.5353 in Table 2. The sampled per-radical values for both models, which are the ones behind Table 3 and Figure 2, ship with the released artifacts (`{model}_rad_cohesions.npy`). The trend is visible but not clean: radicals with transparent, coherent meaning domains cluster at the top, while the bottom mixes positional components such as 刂 (knife) and 彳 (step) with radicals that do name concrete things, 禾 (grain), 车 (cart) and 戈 (halberd). 力 (power) is abstract yet ranks sixteenth. The gradient still points to the radical-aligned structure riding on semantic coherence rather than radical identity as such, but it is a gradient and not a dichotomy.

Rank	Rad.	Meaning	Size	Mean Sim.
1	鸟	bird	66	0.6406
2	足	foot	77	0.6116
3	鱼	fish	55	0.6036
4	疒	sickness	88	0.6021
5	玉	jade	105	0.5874
6	雨	rain	38	0.5872
7	口	mouth	355	0.5851
8	穴	cave	35	0.5797
9	火	fire	135	0.5782
10	羽	feather	22	0.5780
11	马	horse	76	0.5780
12	虫	insect	116	0.5769
13	米	rice	42	0.5741
14	食	food	66	0.5733
15	手	hand	321	0.5706
16	力	power	41	0.5680
17	子	child	29	0.5649
18	金	metal	237	0.5642
19	衤	clothing	73	0.5623
20	牛	ox	24	0.5609

Table 6: Top 20 radicals in mBERT ranked by mean intra-radical cosine similarity (cohesion), computed over all intra-radical pairs.

Rank	Rad.	Meaning	Size	Mean Sim.
59	刂	knife	70	0.5196
60	尸	corpse	32	0.5177
61	戈	halberd	24	0.5174
62	厂	cliff	21	0.5158
63	冫	ice	26	0.5152
64	禾	grain	57	0.5139
65	攴	rap/strike	34	0.5121
66	车	cart	67	0.5109
67	彳	step	39	0.5109
68	广	shelter	47	0.5031

Table 7: Bottom 10 radicals in mBERT ranked by mean intra-radical cosine similarity (cohesion), computed over all intra-radical pairs.

C Semantic Field Similarity Matrices (mBERT)

For each of the four fields we give the pairwise cosine block means from mBERT, computed on

embeddings extracted for the stimulus characters with the procedure of Section 4.1. Characters that carry the target radical come first, the rest after; the comparison that matters is the same-radical block (upper-left) against the cross-radical block. All four fields use the full ten characters, including the two Animals stimuli (虎, 羊) that lie outside the 6,306-character corpus because their own radicals have fewer than 20 members.

Animals field (10×10). Same-radical (犭): 狗, 狼, 狐, 猫, 狮; cross-radical: 虎, 熊, 牛, 羊, 马. Block means: 0.7441 same-radical versus 0.7429 cross-radical, a tiny $\Delta = +0.0012$.

Water field (10×10). Same-radical (氵): 河, 湖, 洋, 泪, 洗; cross-radical: 水, 冰, 雨, 霜, 泉. Here Δ is negative (same-radical mean 0.5894 against cross-radical 0.6202, $\Delta = -0.0308$): inside the 氵 group, semantically far-apart pairs such as 河 (river) and 泪 (tears) drag the intra mean down, while cross-radical pairs like 水/冰 (water/ice) are semantically very close. Sharing a radical is no guarantee of shared meaning.

Wood field (10×10). Same-radical (木): 林, 松, 柳, 桥, 棉; cross-radical: 竹, 藤, 草, 芦, 茎. The negative Δ (-0.0419; same-radical 0.6226, cross-radical 0.6644) comes from semantically tight cross-radical pairs (松/藤, "pine/vine") clustering despite different radicals, while some same-radical pairs (林/桥, "forest/bridge") sit far apart in meaning.

Metal field (10×10). Same-radical (钅): 铁, 铜, 锡, 钢, 银; cross-radical: 金, 玉, 石, 矿, 宝. Block means: 0.6920 same-radical versus 0.6637 cross-radical, $\Delta = +0.0283$.

D Statistical Summary Tables

Permutation test summary. Table 8 collects the permutation statistics. At corpus scale the observed Δ beats almost every shuffle, so the p-values bottom out at the floor; under semantic control Δ sits squarely in the null, and the p-values are non-significant.

Bootstrap intervals. Table 9 gives the bootstrap Δ statistics over 1,000 resamples. Every distribution is symmetric and unimodal, and every 95% interval clears zero.

Holm–Bonferroni correction. Table 10 applies the correction to the four primary tests; all four clear the threshold at $\alpha = 0.05$ with room to spare.

Test	Model/Metric	Obs. Δ	Null SD	p_π
Corpus	mBERT Cos	$+0.0163$	0.0028	9.99×10^{-4}
Corpus	mBERT Euc	$+0.0075$	0.0031	0.0090
Corpus	Ch.-BERT Cos	$+0.0070$	0.0019	0.0020
Corpus	Ch.-BERT Euc	$+0.0052$	0.0011	9.99×10^{-4}
Sem. ctrl	mBERT Cos	-0.0108	0.0169	0.7451
Sem. ctrl	Ch.-BERT Cos	-0.0047	0.0088	0.7089

Table 8: Permutation test summary statistics. Corpus-scale tests use 1,000 shuffles; semantic control tests use 5,000 shuffles.

Model	Metric	Obs. Δ	2.5%	97.5%
mBERT	Cosine	0.0163	0.0106	0.0217
mBERT	Euclidean	0.0075	0.0017	0.0135
Chinese-BERT	Cosine	0.0070	0.0033	0.0106
Chinese-BERT	Euclidean	0.0052	0.0030	0.0073

Table 9: Bootstrap confidence interval statistics for Δ (1,000 resamples). All 95% CIs exclude zero.

Model	Metric	Raw p	Thresh.
mBERT	Cosine	1.65×10^{-8}	0.0125
Chinese-BERT	Euclidean	2.37×10^{-6}	0.0167
Chinese-BERT	Cosine	2.51×10^{-4}	0.0250
mBERT	Euclidean	0.0181	0.0500

Table 10: Holm–Bonferroni multiple-comparison correction applied to the four primary statistical tests. All are significant.

Effect size benchmarks. At corpus scale Cohen's d runs from about 0.057 to 0.137. By the usual yardstick (Cohen, 1988), 0.20 small, 0.50 medium, 0.80 large, these sit below even "small," in what we call a "very small but reliable" band, and the control conditions land at $d \approx 0$. The effects are genuine (they recur across models and metrics and pass every robustness check) yet modest: meaning is the dominant axis of the space, and radical alignment is a faint ripple on top of the semantic regularities the script encodes.

Reproducibility. Numerical outputs are stored as `.npy` and `.csv` files under `results/` and the figures as 300+ dpi PNGs under `figures/`, with the seed fixed at `np.random.seed(42)`. We use `bert-base-multilingual-cased` and `hfl/chinese-bert-wwm-ext` from HuggingFace, with `transformers` $\geq$ 4.20, `scikit-learn` $\geq$ 1.0, `scipy` $\geq$ 1.7, and `numpy` $\geq$ 1.21. Cross-platform floating-point differences may move reproduced values at the fourth decimal, but the qualitative conclusions and significance calls are unaffected. Code, notebooks with inline result images, and

precomputed artifacts are publicly available at: https://github.com/aryan35790jp/chinese-radical-geometry-rocling-38.