Do Compounds and Phrases Follow the Same Combinatorial Principles? What a Multi-Dimensional Probing Reveals for Chinese

Anonymous ACL submission

Abstract

Are compounds and phrases constructed by the same combinatorial principles? This is a long-standing topic in linguistic theory. As a highly isolating language, Chinese offers a particularly revealing case: its word–phrase boundary is weakly marked, thus compound-internal relations often parallel phrase-level syntactic relations. Although this continuity is well described in Chinese linguistics, no operational test has yet examined whether the same principles are similarly represented across the two levels. PLMs provide a shared representational space for operationalizing this question. Motivated by this idea, we build **CompPhrase-7**, a parallel Chinese compound–phrase dataset covering seven shared relation types, and conduct a multi-dimensional probing study of Chinese-language PLMs, which separates three levels of cross-scale continuity: transferable feature information, aligned category geometry, and localized dimension-level causal support. The results show a clear dissociation: feature transfer is stable across models; category geometry depends strongly on pretraining regime; and localized causal support is not detected. The findings point to a differentiated view of Chinese combinatorial principles: compounds and phrases retain comparable relational cues, while lexicalization reshapes how those cues are organized once phrase-like structures become words. This study connects a long-standing linguistic question to a more precise computational representation, showing that continuity can surface as transferable information without shared geometry or localized causal implementation.

1 Introduction

Are compounds and phrases constructed by the same combinatorial principles? This question has long divided linguistic theory: Under the *Lexicalist Hypothesis*, word formation belongs to the lexicon, so compound-internal structure is generally opaque to syntax (Chomsky, 1970); *Distributed Morphology*, by contrast, treats compound-internal structure as syntactically derived, making it possible to analyze compounds and phrases as products of a shared combinatorial system (Halle and Marantz, 1993). At stake in this theoretical contrast is the grammatical relation between word formation and phrase formation: whether compounds and phrases belong to separate domains or represent different scales of a shared combinatorial system.

As a highly isolating language, Chinese offers a particularly revealing case due to its weak morphological marking of the compound–phrase boundary (Li and Thompson, 1981; Packard, 2000). Because Chinese marks grammatical relations primarily through constituent order rather than morphology, the same structural patterns can organize both compounds and phrases. Evidence for this continuity comes from both synchronic and diachronic observations: synchronically, compounds and phrases instantiate relations such as *subject–predicate*, *verb–object*, and *modifier–head* (Zhu, 1982; Shen, 1997; Lu, 2019); diachronically, phrases can gradually lexicalize into compounds through repeated use (Dong, 2002; Shen, 2020). This literature gives strong reasons to expect compound–phrase continuity in Chinese, yet it leaves open whether these parallels can be detected in computational representations.

Thus, Pretrained language models (PLMs) provide such a representational space: compounds and phrases can be encoded by the same model and compared with the same set of probing tools (Belinkov et al., 2017; Hewitt and Manning, 2019). Existing PLM studies, however, have mostly examined morphology and syntax within separate scales, asking whether each type of structure can be decoded from model representations (e.g., Belinkov et al., 2017; Hewitt and Manning, 2019; Hofmann et al., 2021). What remains missing is a cross-scale operationalization for Chinese: a way

to test whether the relation types described in the linguistic literature correspond when they occur inside compounds and across phrase constituents, and to distinguish different kinds of evidence for such correspondence in model representations.

We are motivated to operationalize this comparison with a multi-dimensional probing framework for morphology–syntax isomorphism. The framework asks which forms of representational continuity link Chinese compounds and phrases, and separates three claims that are often conflated in probing work: whether relation-type information is decodable across scales, whether category organization is geometrically aligned, and whether the continuity has localized causal support. It distinguishes three progressively stronger dimensions: *feature isomorphism* (D1), tested by zero-shot cross-scale transfer of linear probes (Tenney et al., 2019); *relational isomorphism* (D2), tested by Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) together with Procrustes alignment (Lample et al., 2018); and *circuit isomorphism* (D3), tested by causal interventions on source-task-critical neural dimensions with matched random controls (Vig et al., 2020; Geiger et al., 2021).

For implementation, we construct CompPhrase-7, a parallel Chinese compound–phrase dataset covering seven construction categories, and apply the multi-dimensional probing to four Chinese PLMs. The results differ across the three dimensions: At D1, all models achieve above-chance cross-scale transfer, indicating shared feature information across compounds and phrases; At D2, cross-scale geometric alignment is present but highly sensitive to the pre-training regime, with the strongest alignment appearing in the whole-word-masking variants; At D3, targeted interventions are no more disruptive than matched random baselines, providing no evidence for localized causal support.

Taken together, the three diagnostics locate compound–phrase continuity at different representational grains: Chinese compounds and phrases preserve relation-type information that can transfer across scales, but lexicalization also reshapes how these relations are organized once they become words. The absence of localized causal support further suggests that this continuity is not concentrated in a small shared set of dimensions. Read in this way, the results move the Chinese compound–phrase debate away from a binary choice between separation and identity. They instead raise a more differentiated question: which aspects of phrase-like structure remain recoverable after word formation, which aspects are reorganized in lexicalized compounds, and which representational levels make these changes visible.

Leveraging this multi-dimensional approach, our contributions are threefold:

(1) A Cross-scale probing framework: We propose a multi-dimensional probing framework for evaluating compound–phrase isomorphism in model representations. By separating feature transfer, category geometry, and localized causal support, the framework prevents decodability, representational organization, and causal evidence from being collapsed into a single notion of similarity;

(2) Linguistic and computational findings: Across Chinese-language PLMs, feature transfer, category geometry, and localized causal support dissociate clearly. It suggests that compounds and phrases can retain comparable relational cues while lexicalization reshapes their organization, and that cross-scale continuity can surface as transferable information without shared geometry or localized causal implementation;

(3) Cross-scale resource: We also contribute **CompPhrase-7**, a parallel Chinese compound–phrase dataset covering seven construction categories, and derive a balanced evaluation subset for controlled cross-scale probing in the future[1].

2 Related Work

2.1 Study on Continuity between Chinese Compounds and Phrases

The relation between morphology and syntax has long been an important point of debate in linguistic theory: The *Lexicalist Hypothesis* posits a strict boundary between the two, treating word formation as an encapsulated lexical process opaque to syntax (Chomsky, 1970; Di Sciullo and Williams, 1987). In contrast, *Distributed Morphology* places word formation within the syntactic derivation, treating morphology and syntax as parts of a shared combinatorial system (Halle and Marantz, 1993; Marantz, 1997; Harley and Noyer, 1999; Embick and Noyer, 2007).

As a highly isolating language, Chinese has played a special role in this debate due to its little inflectional morphology, where compounding

[1]The resources and code will be released to the public once the paper is accepted.

is a major and productive mode of word formation (Li and Thompson, 1981; Dai, 1998; Packard, 2000). These properties make the word–phrase boundary especially difficult to draw cleanly. A central observation in this literature is that Chinese compounds and phrases can instantiate the same structural relation types, including *subject–predicate*, *verb–object*, and *modifier–head* (Zhu, 1982). Later work developed this observation into a broader account of the Chinese word–phrase relation (Shen, 1997; Lu, 2019). This line of work supports structural continuity between compounding and phrasal syntax at the level of constructional relations.

The observation also sharpens a wordhood problem: if a sequence has phrase-like internal organization, when should it nevertheless be analyzed as a word? Work on Chinese compounds has addressed this problem through diagnostics such as phonological constraints (Duanmu, 2000), lexical integrity (Huang, 1984; Dai, 1998; Packard, 2000), semantic integration (Li and Thompson, 1981), and processing evidence (Hoosain, 1992; Cui et al., 2013; Shen et al., 2018). Diachronic accounts connect these perspectives by showing how phrasal patterns can lexicalize into compounds over time (Dong, 2002; Shen, 2020). The resulting picture is continuity under lexicalization: shared constructional relations may persist from phrases to compounds, while lexicalization can change their phonological, grammatical, and semantic behavior (Chi and Lin, 2019). This is the linguistic issue our study mean to operationalizes: we ask how far such continuity survives in PLM representations, distinguishing transferable category information, cross-scale category geometry, and localized causal support.

On the resource side, the computational basis for operationalizing this question has also become available: for word formation, recent datasets make Chinese word-formation categories directly analyzable in computational settings (Zheng et al., 2021); for phrase syntax, the Penn Chinese TreeBank (CTB) provides detailed phrase-structure annotations (Xue et al., 2005). Together, these resources make it possible to construct a controlled comparison between phrase-level syntactic relations and compound-internal relations.

On the model side, representation-analysis studies have examined whether Chinese PLMs encode distinctions relevant to morphology and syntax. Existing studies have found word-boundary and word-structure information (Wang et al., 2020), grammatical knowledge in minimal-pair evaluations (Xiang et al., 2021), and syntactic knowledge in hidden states and attention heads (Zheng and Liu, 2023). Work on segmentation, whole-word masking, and linguistically enriched pretraining further shows that tokenization and masking choices affect how Chinese words and compounds are represented (Dai et al., 2022; Li et al., 2022; Cui et al., 2021; Diao et al., 2020; Sun et al., 2021). These studies show that PLMs encode information relevant to both compounding and syntax. They mostly examine the two levels separately, however, leaving open whether relation types observed inside compounds and across phrase constituents form corresponding structures within the same model.

2.2 Analysis on Language Models: Feature Probing, Representational Similarity, and Causal Analysis

Several lines of work offer complementary ways to analyze what language-model representations encode and how they support model behavior.

Feature probing tests whether target linguistic properties can be decoded from neural representations (Belinkov et al., 2017; Conneau et al., 2018). The probing literature has also emphasized the limits of this evidence: successful decoding shows that information is accessible to a readout, but it does not by itself establish how that information is organized or whether it is used in the model's behavior (Hewitt and Liang, 2019; Voita and Titov, 2020; Elazar et al., 2021).

Representational similarity analysis (RSA; Kriegeskorte et al., 2008) addresses a different question: whether two representation spaces share a similar relational geometry. RSA compares patterns of pairwise dissimilarity, making it possible to ask whether two sets of representations organize their categories or stimuli in similar ways. It has been used to study linguistic dependency structure (Lepori and McCoy, 2020), cross-lingual embedding alignment (Schuster et al., 2019), and layer-wise representation comparison (Abnar and Zuidema, 2020).

Causal interventions ask whether specific internal components causally support model behavior. Causal mediation analysis is used to trace bias pathways in PLMs (Vig et al., 2020), while causal abstraction is formalized as a foundation for mechanistic interpretability (Geiger et al.,

2021, 2025). Related work has extended causal probing to morpho-syntactic phenomena (Amini et al., 2023) and compositional generalization (Kumon and Yanaka, 2025). These methods provide stronger evidence about behavioral relevance than decodability alone, although their interpretation depends on the granularity of the intervention.

Together, these approaches motivate a more differentiated view of representational continuity. Prior Chinese PLM studies establish that information about morphology and syntax is present in model representations, yet they leave open whether relation types observed inside compounds and across phrase constituents are aligned within the same representational space. Our study aims to bring feature probing, similarity analysis, and causal intervention to this cross-scale question.

3 Methodology

Building on the linguistic motivation above, we treat Chinese compounds and phrases as two scales at which the same constructional relation types may appear. We use the multi-dimensional probing framework to test this continuity at three representational levels in Chinese PLMs, summarized in Figure 1: Feature Isomorphism (D1) tests whether category distinctions learned at one scale transfer to the other; Relational Isomorphism (D2) tests whether the two scales preserve similar category geometry; and Circuit Isomorphism (D3) tests for localized causal support.

3.1 Models and Datasets

We evaluate four Chinese PLMs: BERT (`bert-base-chinese`; Devlin et al., 2019), BERT-wwm (`hfl/chinese-bert-wwm-ext`; Cui et al., 2021), ERNIE (`nghuyong/ernie-1.0-base-zh`; Sun et al., 2020), and RoBERTa (`hfl/chinese-roberta-wwm-ext`; Liu et al., 2019; Cui et al., 2021). Full architectural specifications and pre-training configurations are provided in Appendix A. These models are selected to test the generalizability of our findings across different pretraining regimes: The BERT vs. BERT-wwm comparison contrasts a standard Chinese BERT checkpoint with a whole-word-masking variant; while ERNIE and RoBERTa serve as out-of-distribution checks to verify that our findings are not specific to a single model family or training configuration.

CompPhrase-7 covers seven representative construction categories at both word-level compounds and syntax-level phrases, as illustrated in Table 1. The word-level labels are drawn from Zheng et al. (2021), who annotated words with 16 word-formation categories in Chinese. From these categories, we excluded those that are structurally ambiguous, retaining seven categories with clear compound–phrase correspondences. The full list of the original categories and our exclusion criteria are provided in Appendix B. The syntax-level labels are derived from CTB (Xue et al., 2005). It provides detailed phrase-structure annotations where our target construction categories are implicitly encoded within syntactic configurations and functional tags. We extracted phrase constituents from the parse trees and mapped these CTB cues to the seven construction categories using expert-defined criteria. The mapping criteria and verification procedure are described in Appendix B.

CompPhrase-7 is organized in two versions: The full dataset contains 107,963 word-level samples and 193,270 syntax-level samples, but the class distribution is highly skewed. We therefore construct a balanced subset by downsampling each construction category to the same count within each scale: The balanced word-level splits contain 2,611/336/371 train/validation/test samples, and the balanced syntax-level splits contain 10,682/1,470/1,204 samples. All reported analyses use this balanced subset. Detailed mapping criteria, split statistics, and per-category distributions are provided in Appendix B.

3.2 D1: Feature Isomorphism

This dimension tests whether the features that distinguish word-level construction categories also apply to syntax-level construction categories. If a linear classifier trained on word-level representations achieves zero-shot transfer to syntax-level classification, it indicates that both scales rely on a shared set of discriminative features. This constitutes the weakest standard of isomorphism: two scales might be separable by an identical decision boundary even if their category geometry or internal computational mechanisms differ.

Implementation. Word-level compounds and syntax-level phrases are encoded as standalone input sequences. For every model, we extract the [CLS] representation at the shared top encoder layer (L12) and train a 7-way linear probe on one scale, then evaluate zero-shot transfer on the other

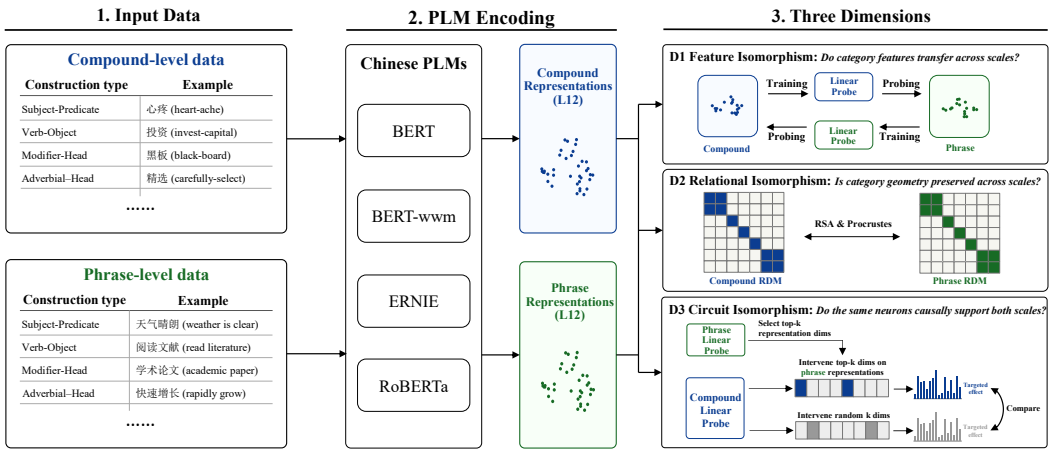

Figure 1: Overview of the multi-dimensional probing framework. Using the same Chinese PLMs and aligned word- and syntax-level construction categories, D1 tests transferable features with linear probes, D2 tests relational geometry with RSA and Procrustes analysis, and D3 tests localized causal support with same-layer interventions against matched random controls.

Table 1: The seven construction categories in CompPhrase-7, demonstrating the structural parallelism between word-level compounds and syntax-level phrases.

Category	Chinese	Word-Level	Phrase-Level
subject–predicate	主谓	心疼 (heart-ache)	天气晴朗 (weather is clear)
verb–object	述宾	投资 (invest-capital)	阅读文献 (read literature)
modifier–head	定中	黑板 (black-board)	学术论文 (academic paper)
adverbial–head	状中	精选 (carefully-select)	快速增长 (rapidly grow)
verb–complement	述补	改善 (improve-perfect)	打扫干净 (clean thoroughly)
serial–verb	连谓	拜访 (visit-consult)	去图书馆看书 (go to library to read)
preposition–object	介宾	于是 (thus-at)	在教室里 (in the classroom)

(w→s and s→w). We also report within-scale references (w→w and s→s). Appendix C provides the full probe specification, per-category transfer analysis, and a mean-pooling robustness check over non-special token representations.

3.3 D2: Relational Isomorphism

This dimension measures whether the relative distances among construction categories are preserved across scales. For instance, if *verb–object* and *verb–complement* categories are nearest neighbors in the word-level representation space, do they remain nearest neighbors in the syntax-level space? This is a medium-strength standard: it asks whether the two scales share a category-level geometry, beyond the feature separability in D1.

Implementation. For each model and layer, we average the [CLS] representations of all samples within each construction category to obtain seven word-level prototypes and seven syntax-level prototypes. We then compute a 7×7 Representational Dissimilarity Matrix (RDM) for each scale, where each entry is the Euclidean distance between two category prototypes. RSA measures whether the two RDMs rank category-pair distances similarly, using Spearman correlation over the 21 off-diagonal distance pairs. As a complementary check, orthogonal Procrustes analysis tests whether the word-level prototype configuration can be rotated into the syntax-level configuration with low disparity. Full statistical procedures, including bootstrap confidence intervals for RSA and permutation testing for Procrustes alignment, are provided in Appendix D.

3.4 D3: Circuit Isomorphism

This dimension tests a narrower and stronger claim than D1 and D2: whether the same small set of representation dimensions that supports probe decisions at one scale also has causal leverage over probe decisions at the other scale. In this sparse dimension-level formulation, cross-scale circuit isomorphism predicts that intervening on word-critical dimensions should change syntax predictions, and intervening on syntax-critical dimensions should change word predictions, more than

Table 2: Bidirectional zero-shot cross-scale transfer with L12 linear probes. Target test sets contain 1,204 syntax-level samples for w→s and 371 word-level samples for s→w; both are balanced over seven classes, giving identical 14.3% random and majority baselines.

Model	**w→s Acc. [95% CI]**	**s→w Acc. [95% CI]**
BERT	30.2 [0.277, 0.328]	28.8 [0.245, 0.337]
BERT-wwm	33.1 [0.305, 0.358]	27.7 [0.234, 0.325]
ERNIE	25.1 [0.227, 0.276]	28.9 [0.245, 0.337]
RoBERTa	24.9 [0.225, 0.274]	25.1 [0.210, 0.298]
Random / Majority	14.3	14.3

interventions on matched random dimensions. D3 therefore evaluates localized causal support for cross-scale continuity.

Implementation. We evaluate D3 bidirectionally at L12, so dimension selection and intervention occur in the same representational space. For each direction, we use the D1 task-specific linear probe to rank L12 dimensions by gradient-based importance for source-scale classification, select the Top-50 source-critical dimensions, and intervene on opposite-scale target representations: word-critical dimensions for syntax-level inputs in w→s, and syntax-critical dimensions for word-level inputs in s→w. The main intervention suppresses the selected dimensions and measures changes in the target-scale probe's predicted probabilities; a robustness check instead replaces them with corresponding source-scale category prototype values. To test specificity, we repeat the same procedure on 100 equally sized random dimension sets and report the targeted-vs.-random fold change. Appendix E gives full details on dimension selection, intervention mechanics, and statistical testing.

4 Results

4.1 Dimension 1: Feature Isomorphism

D1 asks whether category distinctions learned at one scale transfer to the other. Table 2 reports the balanced-data zero-shot transfer results with linear probes at L12. All four models remain well above the 14.3% baseline in both directions: w→s accuracy ranges from 24.9% to 33.1%, and s→w accuracy ranges from 25.1% to 28.9%. This bidirectional pattern provides consistent evidence for cross-scale feature transfer across models.

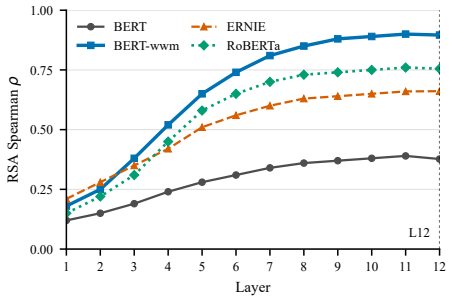

Figure 2: Layer-wise RSA between word- and syntax-level category RDMs. Alignment increases in higher layers, especially under whole-word-masking pretraining; the dashed line marks L12 values reported in Table 3.

Table 3: RSA and Procrustes alignment results in L12. Both metrics indicate strong alignment for whole-word-masking variants, though their exact rankings differ.

Model	**RSA ρ**	**95% CI**	**Proc.**	**Perm. p**
BERT	0.377	[−0.11, 0.74]	0.232	0.056
BERT-wwm	**0.896**	[0.74, 0.96]	**0.088**	< 0.001
ERNIE	0.661	[0.32, 0.86]	0.177	0.028
RoBERTa	0.755	[0.44, 0.92]	0.077	< 0.001

The absolute scores are modest, yet the task is demanding: a linear classifier trained at one scale must generalize directly to another scale with different granularity, distribution, and context structure. In this setting, 25–33% accuracy shows that some construction-discriminative regularities survive cross-scale transfer without target-scale training. The within-scale references in Appendix C.2.1 are much higher (62.6–69.0% for w→w and 67.0–75.7% for s→s), which places the D1 result in a partial-sharing range. The two scales share transferable category information, while much of the scale-specific structure remains easier to recover within each scale. A per-category analysis in Appendix C.2.2 further shows that transfer varies across construction types. The uneven pattern points to construction-specific regularities as the main source of the transfer signal.

4.2 Dimension 2: Relational Isomorphism

D2 asks whether the shared information observed in D1 is also organized similarly across scales. Figure 2 shows that RSA increases with depth across all four models, but the trajectories diverge sharply: BERT remains weak throughout, whereas BERT-

Table 4: Bidirectional same-layer suppress-intervention results at L12. Target sets contain 1,204 syntax samples for w→s and 371 word samples for s→w. "Rand." denotes the matched random-dimension control baseline.

Model	w→s			s→w		
	Int.	Rand.	Fold	Int.	Rand.	Fold
BERT	69.1%	69.7%	0.99×	71.4%	72.4%	0.99×
BERT-wwm	81.1%	78.1%	1.04×	89.8%	90.3%	0.99×
ERNIE	89.0%	85.7%	1.04×	96.5%	96.3%	1.00×
RoBERTa	90.0%	89.9%	1.00×	93.5%	92.7%	1.01×

wwm rises rapidly after the middle layers and reaches near-perfect alignment by L12. Table 3 confirms the same pattern with the L12 statistics. The key contrast is BERT vs. BERT-wwm: the whole-word-masking BERT variant raises RSA from 0.377 to 0.896 and lowers Procrustes disparity from 0.232 to 0.088.

Table 3 and Figure 2 therefore separate D1-style feature transfer from D2-style geometric alignment. BERT shows above-baseline cross-scale transfer in D1, but its category geometry remains only weakly aligned across scales. BERT-wwm shows a stronger pattern: its word-level and syntax-level category geometries become closely aligned in the upper layers. This contrast suggests that relational isomorphism is sensitive to pre-training regime, with whole-word masking producing the clearest cross-scale geometric alignment.

The BERT result should still be read cautiously. Its RSA correlation is not statistically significant ($\rho = 0.377$, $p = 0.114$), but D2 has limited statistical power with only $N = 7$ categories, and the corresponding Procrustes result is marginal ($p = 0.056$). We therefore treat BERT as a weak-alignment case. Appendix D provides the full procedure, with layer-wise RSA and Procrustes results in Appendix D.2.1 and the power analysis in Appendix D.2.2. D2 thus supports relational isomorphism as a training-sensitive property, while leaving open whether this alignment is supported by the same localized dimensions.

4.3 Dimension 3: Circuit Isomorphism

D3 asks whether the cross-scale signal identified by D1 and D2 has localized causal support. The diagnostic criterion is specificity: dimensions selected as important for one scale should affect opposite-scale predictions more than matched random dimensions of the same size.

Table 4 reports the bidirectional same-layer suppress-intervention results. Each direction suppresses dimensions critical to the source scale and evaluates predictions at the opposite scale. The interventions are strong in absolute terms, affecting 69.1–90.0% of samples in the w→s direction and 71.4–96.5% in the s→w direction. The key result is the comparison with matched random controls: across all four models and both directions, the targeted-vs.-random fold change remains effectively 1.0×. The selected source-critical dimensions therefore show no cross-scale specificity.

The same targeted-vs.-random pattern holds under smaller intervention sets (Appendix E.2.2). D3 therefore narrows the interpretation of D1 and D2: the cross-scale signal is readable, and in some models geometrically aligned, while the selected localized dimensions show no specificity over matched random controls.

5 Discussion

5.1 Interpretation of Theoretical Aspects

The results point to a differentiated account of Chinese compound–phrase continuity. The evidence is strongest for transferable category distinctions, conditional for shared category geometry, and absent for shared localized causal support. This pattern shifts the theoretical focus away from a binary compound–phrase opposition toward the specific properties that remain continuous, reorganized, or weakened when phrase-like relations become lexical words.

D1 supports the weaker side of this view. Chinese compounds and phrases share enough relational structure for the model to reuse construction-discriminative cues across scales, which is consistent with the long-standing observation that the two levels can instantiate parallel relation types. This is continuity at the level of transferable category distinctions: the relation is still readable, but only in that limited sense.

D2 and D3 define the boundary of that continuity: The training-sensitive geometry in D2 suggests that shared relation types are reorganized once phrase-like structures are lexicalized as words; D3 then shows that this continuity does not yield localized causal support. Taken together, the results point to a differentiated view of the Chinese compound–phrase relation: compounds and phrases can share constructional relations, while lexicalization can alter how those relations are represented once they become words. For Chinese lin-

guistic theory, the issue is therefore which properties remain continuous across scales and which are reorganized by word formation.

5.2 Why the Diagnostics Diverge

The divergence among D1, D2, and D3 is expected if the three diagnostics depend on different representational conditions.

D1 requires only reusable relation-discriminative cues across scales. In Chinese, such cues can arise because the same structural relations can be realized both inside compounds and across phrase constituents, including *modifier–head* order, predicate–argument relations, *verb–complement* relations, and *serial–verb* relations (Zhu, 1982; Shen, 1997; Lu, 2019; Dong, 2002; Shen, 2020). A PLM can therefore learn cues that remain useful across scales even when one instance appears inside a lexicalized word and the other appears in a larger syntactic combination. Such transfer does not require the full organization of word-level and syntax-level categories to match.

D2 imposes a stricter condition. Cross-scale geometry can align only when the model organizes construction categories in comparable ways at the two scales. Character- or subword-level encoding can preserve enough information for category discrimination while still distributing that information across a compound's internal components. Whole-word-masking pre-training places more pressure on the model to integrate a compound before comparing it with other categories. This pressure fits the Chinese case: many compounds have analyzable internal relations, yet they function as lexical words (Huang, 1984; Dai, 1998; Packard, 2000). Psycholinguistic evidence points in the same direction, suggesting that constituent characters remain informative during Chinese compound processing, but successful comprehension depends on integrating them into a whole-compound representation (Cui et al., 2013; Shen et al., 2018). The D2 pattern therefore points to wordhood and lexical integrity as likely determinants of cross-scale geometry.

D3 requires a more specific condition: the cross-scale signal must have localized causal support. Inside lexicalized compounds, shared relation types are embedded in word-level representations shaped by the compound's status as an integrated lexical item (Huang, 1984; Dai, 1998; Packard, 2000; Chi and Lin, 2019). This makes it plausible for construction categories to remain identifiable across scales, and geometrically aligned in some models, while their implementation becomes more distributed after lexicalization (Dong, 2002; Shen, 2020). Eye-movement evidence is consistent with this distinction: lexicalized and novel Chinese compounds recruit semantic information differently during reading (Wang et al., 2023). D3 completes this stratified picture: the cross-scale signal that D1 makes readable and D2 partly aligns geometrically does not extend to localized causal support, consistent with shared relations being re-embedded in word-specific representations after lexicalization.

6 Conclusion

Our multi-dimensional probing framework operationalizes Chinese compound–phrase continuity as a cross-scale representational question, separating feature transfer, category geometry, and localized causal support — claims often collapsed into a single notion of representational similarity. Across Chinese PLMs, the diagnostics yield a stratified pattern: stable cross-scale feature transfer, training-sensitive category geometry, and no detected localized causal support. It points to a differentiated view of Chinese combinatorial principles: Compounds and phrases retain comparable relation-type information, but lexicalization can reshape how those relations are organized once phrase-like structures become words. The findings therefore connect a long-standing linguistic question to a more precise computational representation: continuity may be visible in what a model makes decodable, without requiring the same geometric organization or the same localized causal implementation.

Future work can be further explored in such directions: Cross-linguistic studies can test whether the same stratified pattern holds in languages with richer inflection, clearer word boundaries, or different compounding systems; Model comparisons can further isolate how segmentation, whole-word masking, and pre-training objectives affect cross-scale geometry; More fine-grained interventions can ask when compound–phrase continuity remains distributed across representational patterns and when it becomes causally localized.

Limitations

The main limitations concern the scope of the linguistic materials, the model family tested, the input

conditions of context and cue, and the granularity of the causal interventions:

About linguistic scope. The study focuses on Chinese and on seven construction categories with clear word–phrase correspondences. This design is appropriate for testing cross-scale comparability, but it leaves open whether the same layered pattern would hold for less canonical categories in Chinese or for languages with richer morphology and a less permeable word–phrase boundary;

About model scope. We evaluate four Chinese encoder-only PLMs pretrained on Chinese corpora, all within the BERT family. The findings are therefore most directly grounded in Chinese BERT-family encoders, and broader generalization to decoder-only PLMs, multilingual models, or larger architectures requires further testing. Because these checkpoints differ in several pretraining and segmentation choices, model-level contrasts should be interpreted as architecture- and training-regime effects rather than as isolated effects of any single design choice;

Context and cue controls. The main experiments encode compounds and phrases as standalone strings. This keeps the two scales comparable, but it does not test whether the same trends hold when compounds and phrases are embedded in richer sentence contexts. Some categories may also contain overt lexical cues, such as prepositional material in *preposition–object* constructions. Our per-category analysis helps identify where transfer is concentrated, but future work should add masked-cue and templated-context controls;

About intervention scope. D3 tests circuit isomorphism through probe-mediated dimension-level and attention-head interventions. This is strong enough to challenge a sparse, localized shared bottleneck for decisions over construction categories, but it does not exhaust the space of possible causal implementations in the base masked-language model. More distributed subspace-level mechanisms, cross-layer pathways, activation-patching effects, or effects that emerge more clearly in larger models remain open for future work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of ACL*, pages 4190–4197.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics (TACL)*, 11:264–281.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of ACL*, pages 861–872.

Changhai Chi and Zhiyong Lin. 2019. "汉语复合词的结构与句法结构平行" 说新议 [a new discussion on the parallelism between compound word structure and syntactic structure in chinese]. *Journal of Zhejiang University (Humanities and Social Sciences)*, 49(5):210–223. In Chinese.

Noam Chomsky. 1970. Remarks on nominalization. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 184–221. Ginn, Waltham, MA.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, pages 2078–2088.

Lei Cui, Guoli Yan, Xuejun Bai, Jukka Hyönä, Suiping Wang, and Simon P. Liversedge. 2013. Processing of compound-word characters in reading chinese: An eye-movement-contingent display change study. *Quarterly Journal of Experimental Psychology*, 66(3):527–547.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2021. Pre-Training with Whole Word Masking for Chinese BERT. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 29, pages 3504–3514.

John Xiang-Ling Dai. 1998. Syntactic, phonological, and morphological words in Chinese. In Jerome L. Packard, editor, *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, pages 103–134. Mouton de Gruyter, Berlin.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. 2022. "is Whole Word Masking always better for Chinese BERT?" probing on Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Anna Maria Di Sciullo and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge, MA.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese text encoder enhanced by N-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740.

Xiufang Dong. 2002. Cihuihua: Hanyu Shuangyinci de Yansheng he Fazhan *[Lexicalization: The Origin and Evolution of Chinese Disyllabic Words]*. Sichuan Nationalities Press, Chengdu. In Chinese.

San Duanmu. 2000. *The Phonology of Standard Chinese*. Oxford University Press, Oxford.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 9, pages 10–25.

David Embick and Rolf Noyer. 2007. Distributed morphology. In Gillian Ramchand and Charles Reiss, editors, *The Oxford Handbook of Linguistic Interfaces*, pages 289–324. Oxford University Press, Oxford.

Atticus Geiger, Duligur Ibeling, Amir Zur, Mahe Chaudhary, Shreyas Kaushik, William Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26:1–63.

Atticus Geiger, Haoxuan Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstraction of neural networks. *arXiv preprint arXiv:2106.02997*.

Morris Halle and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale and Samuel Jay Keyser, editors, *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, pages 111–176. MIT Press, Cambridge, MA.

Heidi Harley and Rolf Noyer. 1999. Distributed morphology. In *Glot International State-of-the-Article*, volume 3, pages 3–9.

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

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, pages 4129–4138.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of ACL*, pages 3594–3608.

Rumjahn Hoosain. 1992. Psychological reality of the word in Chinese. In Hsuan-Chih Chen and Ovid J. L. Tzeng, editors, *Language Processing in Chinese*, pages 111–130. Elsevier, Amsterdam.

C.-T. James Huang. 1984. Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association*, 19(2):53–78.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Ryoma Kumon and Hitomi Yanaka. 2025. Analyzing the inner workings of transformers in compositional generalization. In *Proceedings of NAACL-HLT*, page 432.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International conference on learning representations*.

Michael Anthony Lepori and R. Thomas McCoy. 2020. Picking BERT's brain: Probing for linguistic dependencies using representational similarity analysis. In *Proceedings of COLING*, pages 3632–3644.

Charles N. Li and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, CA.

Wei Li, Yuhan Song, Qi Su, and Yanqiu Shao. 2022. Unsupervised Chinese word segmentation with BERT oriented probing and transformation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3935–3940.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jianming Lu. 2019. Xiandai Hanyu Yufa Yanjiu Jiaocheng *[A Course in Modern Chinese Grammar Research]*, 5th edition. Peking University Press, Beijing. In Chinese.

Alec Marantz. 1997. No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics*, 4(2):201–225.

Jerome L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press, Cambridge.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT*, pages 1599–1613.

Jiaxuan Shen. 2020. Hanyu Da Yufa Wu Lun *[Five Lectures on Chinese Grammar]*. Shanghai Century Publishing Group, Shanghai. In Chinese.

Wei Shen, Xingshan Li, and Alexander Pollatsek. 2018. The processing of chinese compound words with ambiguous morphemes in sentence context. *Quarterly Journal of Experimental Psychology*, 71(1):131–139.

Yang Shen. 1997. *Xiandai Hanyu Fuheci de Dongtai Leixing* [dynamic types of modern Chinese compounds]. In *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistic Studies]*. Beijing Language and Culture University Press, Beijing. In Chinese.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2020. ERNIE: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

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 and the 11th International Joint Conference on Natural Language Processing*, pages 2065–2075.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and 1 others. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12388–12401.

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

Jingwen Wang, Jinmian Yang, Chris Biemann, and Xingshan Li. 2023. Mechanism of semantic processing of lexicalized and novel compound words: An eye movement study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(11):1812–1822.

Yile Wang, Leyang Cui, and Yue Zhang. 2020. Does Chinese BERT encode word structure? In *Proceedings of COLING*, pages 2826–2836.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2784–2790.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Hua Zheng, Yaqi Yan, Yue Wang, Damai Dai, and Yang Liu. 2021. 基于词信息嵌入的汉语构词结构识别研究 (Chinese word-formation prediction based on representations of word-related features). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 386–397.

Jianyu Zheng and Ying Liu. 2023. What does Chinese BERT learn about syntactic knowledge? *PeerJ Computer Science*, 9:e1478.

Dexi Zhu. 1982. Yufa Jiangyi *[Essays on Grammar]*. Commercial Press, Beijing. In Chinese.

A Model Specifications

All four models use BERT-base encoder architectures with 12 layers, 768 hidden dimensions, 12 attention heads, and a maximum sequence length of 512. Table 5 therefore focuses on the checkpoints used in our experiments and the pre-training differences most relevant to our comparisons.

The BERT vs. BERT-wwm comparison contrasts a standard Chinese BERT checkpoint with a whole-word-masking variant, while ERNIE and RoBERTa extend the comparison to related but non-identical pre-training regimes.

B Construction of CompPhrase-7 and Annotation Guidelines

B.1 Rationale of Category Selection

The word-level morphological labels are drawn from Zheng et al. (2021), who annotate 16 word-formation categories in Chinese. Table 6 lists the full taxonomy used in that dataset; the seven construction categories retained in our experiments are shown in bold.

We retain seven construction categories because they denote binary structural relations that can be mapped consistently to syntax-level categories in CTB: 主谓 **(subject–predicate)**, 述宾 **(verb–object)**, 定中 **(modifier–head)**, 状中 **(adverbial–head)**, 连谓 **(serial–verb)**, 述补 **(verb–complement)**, and 介宾 **(preposition–object)**. The remaining nine are excluded for three reasons: First, some do not provide a comparable binary word–phrase relation for our task, including 单纯 *(Single Morpheme)*, 重叠 *(Overlapping)*, 前缀 *(Prefixation)*, and 后缀 *(Suffixation)*; Second, some lack a stable syntax-level counterpart under our CTB mapping, including 方

Table 5: Pretrained checkpoints and salient pre-training differences for the four models used in this study. Shared architectural settings are summarized in the text.

Model	Checkpoint	Vocab.	Pre-training Distinction
BERT	`bert-base-chinese`	21,128	Standard Chinese BERT with character-level masking
BERT-wwm	`hfl/chinese-bert-wwm-ext`	21,128	Whole-word masking variant of Chinese BERT
ERNIE	`nghuyong/ernie-1.0-base-zh`	18,000	Knowledge-enhanced pre-training with phrase/entity masking
RoBERTa	`hfl/chinese-roberta-wwm-ext`	21,128	Whole-word masking with larger-scale RoBERTa pre-training

Table 6: The taxonomy of 16 Chinese word-formation categories from Zheng et al. (2021). Each category is shown with a brief definition, a representative example, and its lexical prevalence. The seven construction categories used in this study are shown in bold.

Word-Formation Rule	Explanation	Use Case	%
定中 **(modifier–head)**	morph_1 modifies morph_2 (noun).	红花 (red-flower)	38.62
联合 (Parallel)	morph_1 and morph_2 are similar or complementary.	昏花 (dizzy-dim)	22.87
述宾 **(verb–object)**	morph_1 operates on morph_2.	花钱 (spend-money)	16.44
状中 **(adverbial–head)**	morph_1 modifies morph_2 (verb).	白花 (vainly-spend)	8.45
单纯 (Single Morpheme)	The word is a single morpheme.	花生 (peanut)	3.51
连谓 **(serial–verb)**	morph_2 is the consequence of morph_1.	休息 (stop-rest)	3.43
后缀 (Suffixation)	morph_2 is the suffix of morph_1.	花头 (trick-∅)	2.70
述补 **(verb–complement)**	morph_2 is the action follows morph_1.	压低 (press-down)	1.28
主谓 **(subject–predicate)**	morph_1 is the subject of morph_2.	眼花 (eyesight-dim)	1.06
重叠 (Overlapping)	morph_1 and morph_2 are the same.	白白 (vainly-vainly)	0.59
方位 (Entity-Position)	morph_1 is an entity, morph_2 is a position.	期中 (semester-mid)	0.37
介宾 **(preposition–object)**	morph_1 is a preposition, morph_2 is an object.	凭空 (from-nowhere)	0.31
名量 (Noun-Quantifier)	morph_2 is the quantifier of morph_1.	花朵 (flower-bud)	0.13
数量 (Number-Quantifier)	morph_1 is a number, morph_2 is a quantifier.	一点 (one-dot)	0.10
前缀 (Prefixation)	morph_1 is the prefix of morph_2.	老师 (∅-teacher)	0.10
复量 (Quantifier-Quantifier)	Both morph_1 and morph_2 are quantifiers.	千米 (kilo-meter)	0.03

位 *(Entity–Position)*, 名量 *(Noun–Quantifier)*, 数量 *(Number–Quantifier)*, and 复量 *(Quantifier–Quantifier)*; Third, 联合 *(Parallel)* was not retained because the boundary between compounding and coordination is especially ambiguous, making cross-scale mapping unreliable. Concrete word- and phrase-level examples for the seven construction categories are given in Table 1 in the main text.

B.2 Details of Data Construction

Full dataset and balanced subset. CompPhrase-7 is organized in two versions. The full dataset contains the original train/validation/test splits for the seven construction categories. Because these splits are strongly imbalanced, we derive a balanced version by downsampling each category within each scale to a common count. All reported analyses use this balanced subset. D1 uses it to control probe capacity and label frequency when interpreting zero-shot transfer; D2 uses it to compute category prototypes from comparable word- and syntax-level samples; and D3 uses it for source-task importance estimation and target-side intervention evaluation;

Dataset statistics. The full CompPhrase-7 dataset contains 107,963 word-level samples and 193,270 syntax-level samples. The class imbalance is substantial: word-level category counts range from 474 to 63,264 samples, and syntax-level counts range from 1,908 to 75,942 samples. The balanced subset therefore uses the smallest available class at each scale, giving 474 word-level samples per category and 1,908 syntax-level samples per category. This yields 2,611/336/371 train/validation/test samples at the

word level and 10,682/1,470/1,204 at the syntax level. Because the balanced test sets contain seven equal-size classes, both the uniform random and majority-class baselines are $1/7 \approx 14.3\%$. Per-category counts are reported in Table 8;

Word-level samples. Word-level samples were drawn from Zheng et al. (2021). We excluded items whose word-formation analysis was unclear (i.e., the same surface form could receive multiple structural labels) and items that were too rare for reliable analysis, operationalized as words with zero occurrences in the reference corpus used for frequency filtering;

Syntax-level label mapping. Syntax-level labels were derived from the Penn Chinese Treebank (CTB; Xue et al., 2005). CTB does not directly annotate our seven construction categories as labels. We therefore derived them from local phrase-structure configurations and functional tags. Table 7 lists the CTB label combinations treated as positive cues for each category. These combinations are not one-to-one replacements for the construction labels; they define the candidate structures that were then checked against the ambiguity conditions in the rightmost column.

B.3 Sample Distribution

Table 8 gives the per-category full-corpus totals and balanced train/validation/test counts.

C D1: Feature Isomorphism

C.1 Implementation Details

C.1.1 Linear Probe Formulation

Word-level compounds are fed to the tokenizer as standalone strings, and syntax-level phrases are fed as standalone phrase strings. For each model, we extract the [CLS] representation at layer 12 (L12), which provides a shared fixed comparison point across all four encoder models. The probe is a linear probe mapping $\mathbb{R}^{768}$ to the seven construction categories. Given a representation $h_i \in \mathbb{R}^{768}$, the probe first computes linear logits

$$z_i = Wh_i + b, \tag{1}$$

where $W \in \mathbb{R}^{7\times 768}$ and $b \in \mathbb{R}^7$. Class probabilities are then obtained by

$$\hat{y}_i = \mathrm{softmax}(z_i). \tag{2}$$

Training minimizes the standard cross-entropy objective

$$\mathcal{L}_{\mathrm{D1}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{7} y_{i,c}\log \hat{y}_{i,c}. \tag{3}$$

C.1.2 Training Setup

We optimize with Adam (learning rate 10^{-3}, batch size 32) for 10 epochs and repeat each setting with 5 random seeds (42–46).

C.1.3 Confidence Intervals

For all classification accuracy results in D1, we report 95% Wilson Score confidence intervals rather than normal-approximation (Wald) intervals. The Wilson Score interval is preferred for proportion estimates because it maintains proper coverage even when the proportion is near 0 or 1, and it performs well with moderate sample sizes.

The Wilson Score interval is computed as:

$$\mathrm{CI} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{4}$$

where $\hat{p}$ is the observed accuracy, n is the number of test samples, and $z = 1.96$ for 95% confidence. For the cross-scale transfer results ($n = 1{,}204$), the Wilson Score intervals provide reliable uncertainty estimates that account for the finite sample size.

C.2 Supplementary Results

C.2.1 Within-Scale Reference Results

Table 9 reports the within-scale reference results at L12. The two cross-scale directions are given in the main text (Table 2).

Within-scale performance remains much higher than cross-scale transfer, indicating that shared features form only part of the representational basis at each scale.

C.2.2 Per-Category Transfer

Table 10 reports per-category recall for the L12 linear probes used in the main D1 analysis. Cross-scale transfer is uneven: *modifier–head* is consistently strong, and *serial–verb* is strong especially in the s→w direction, whereas *subject–predicate*, *verb–object*, and *verb–complement* vary substantially across models and directions. This unevenness is useful diagnostically. It suggests that D1 is not simply driven by a uniform label prior or by all

Table 7: Operational CTB label combinations used to derive syntax-level construction labels.

Category	Positive CTB label combinations	Ambiguity checks
subject–predicate	IP → NP-SBJ/DP-SBJ/QP-SBJ/IP-SBJ + VP; predicative clauses with NP-PRD/QP-PRD/ ADJP-PRD/IP-PRD.	Topic-comment structures with NP-TPC; headlines or fragments without an explicit predicative relation; cases where the apparent subject is an extracted trace.
verb–object	VP headed by VV/VC/VE with NP-OBJ, NP-PN-OBJ, DP-OBJ, QP-OBJ, IP-OBJ, or CP-OBJ.	Resultative or directional complements labeled as objects; passive/ba constructions where the object relation is not local to the candidate VP.
modifier–head	Nominal projections such as NP → ADJP/NP/QP/ DP/CP/DNP + NP; genitive and relative modifiers such as DNP + DEG + NP and CP + DEC + NP.	Appositive structures (NP-APP); coordination with CC or punctuation; noun–noun sequences whose relation is classificatory or parallel rather than *modifier–head*.
adverbial–head	Predicate projections such as VP → ADVP/DVP/ QP-ADV/NP-ADV + VP; adjunct phrases such as PP-ADV, PP-MNR, PP-TMP, PP-LOC, PP-DIR, NP-TMP, LCP-TMP, and LCP-LOC attached to a following VP.	Clause-level adjuncts whose scope is the whole IP rather than the local predicate; locative or temporal NPs used as arguments; discourse connectives.
verb–complement	Complement configurations within VP, especially verbal heads followed by QP-EXT/NP-EXT/LCP-EXT, predicative complements (ADJP-PRD, VP-PRD), or CTB verbal-complement tags such as VRD, VPT, and VCP.	*serial–verb* event chains; *verb–object* cases where the postverbal phrase is selected as an object; coordinated verbs labeled with VCD.
serial–verb	Consecutive verbal projections such as VP → VP + VP, VP → VV + VP, or coordinated/compound verbal labels such as VCD and VSB when they encode an event chain.	Coordination marked by CC or punctuation; complement structures where the second predicate describes the result/state of the first; control clauses with an independent embedded subject.
preposition–object	Prepositional or coverb phrases such as PP → P + NP/NP-PN/DP/QP/IP/LCP; functional variants include PP-ADV, PP-LOC, PP-TMP, PP-DIR, PP-MNR, PP-BNF, PP-PRP, and PP-LGS.	Coverb uses that participate in *serial–verb* structures; prepositional modifiers embedded inside larger DNP or NP structures; cases where the apparent object is a trace.

categories being equally easy; instead, some construction relations provide more stable cross-scale cues than others.

C.2.3 Pooling Robustness

Because Chinese BERT-family tokenizers represent many compounds as character sequences, we also repeated the balanced D1 experiment with L12 mean-pooled token representations. The pooling excludes [CLS], [SEP], and padding tokens and averages the remaining token states. Table 11 shows that the above-chance cross-scale pattern remains under mean pooling: w→s accuracy ranges from 27.1% to 32.4%, and s→w accuracy ranges from 29.8% to 31.3%. Within-scale accuracy also increases under mean pooling, especially for word-level compounds. Thus, the main D1 conclusion is not an artifact of relying only on [CLS], although the stronger within-scale scores indicate that token-level constituent information remains useful.

D D2: Relational Isomorphism

D.1 Implementation Details

D.1.1 Prototype and RDM Construction

For each model and layer, we use the balanced subset and the [CLS] representations of the seven construction categories. We compute a prototype vector per category by averaging all sample representations belonging to that category. Let S_c denote the set of representations assigned to category c. The prototype for that category is

$$p_c = \frac{1}{|S_c|} \sum_{h_i \in S_c} h_i. \tag{5}$$

Using these prototypes, we construct a 7×7 Representational Dissimilarity Matrix (RDM) at each scale. For categories a and b, the corresponding matrix entry is the Euclidean distance

$$R_{ab} = \|p_a - p_b\|_2. \tag{6}$$

This yields one RDM for the word-level prototypes and one for the syntax-level prototypes at each model layer.

Table 8: Per-category sample counts in the full corpus and in the balanced train/validation/test splits.

Category	Word Full	Sentence Full	Word Balanced (train/val/test)	Sentence Balanced (train/val/test)
subject–predicate	1,339	36,391	373/48/53	1,526/210/172
verb–object	22,250	35,282	373/48/53	1,526/210/172
modifier–head	63,264	75,942	373/48/53	1,526/210/172
adverbial–head	12,712	24,704	373/48/53	1,526/210/172
serial–verb	5,739	7,711	373/48/53	1,526/210/172
verb–complement	2,185	1,908	373/48/53	1,526/210/172
preposition–object	474	11,332	373/48/53	1,526/210/172

Table 9: Within-scale reference accuracies for the balanced-data linear probes at L12.

Model	w→w	s→s
BERT	69.0	73.9
BERT-wwm	64.7	75.6
ERNIE	62.6	73.5
RoBERTa	64.0	67.0

D.1.2 RSA and Bootstrap Estimation

We compute the Spearman rank correlation ρ between the upper-triangular portions of the two RDMs (21 unique distance pairs). If $\mathrm{vec}_\triangle(\cdot)$ denotes vectorization of the upper-triangular entries excluding the diagonal, then

$$\rho = \mathrm{Spearman}\Big(\mathrm{vec}_\triangle(R^{(w)}), \mathrm{vec}_\triangle(R^{(s)})\Big). \quad (7)$$

To assess statistical significance, we perform bootstrap resampling with 10,000 resamples (sampling with replacement from the 21 distance pairs) and report 95% bootstrap confidence intervals.

D.1.3 Procrustes Alignment and Permutation Test

As a complementary geometric check, we perform orthogonal Procrustes alignment (without scaling) between the two prototype sets. Let $P^{(w)}, P^{(s)} \in \mathbb{R}^{7\times d}$ denote the matrices whose rows are the seven word-level and syntax-level prototypes. We solve

$$Q^\star = \arg\min_{Q^\top Q=I} \|P^{(w)}Q - P^{(s)}\|_F, \quad (8)$$

where Q is an orthogonal transformation and $\|\cdot\|_F$ is the Frobenius norm. We then compute the Procrustes similarity score from the aligned prototype sets and assess significance via permutation testing over all $7! = 5{,}040$ possible label permutations. A p-value is computed as the proportion of random permutations yielding a Procrustes score equal to or greater than the observed alignment.

D.2 Supplementary Results

D.2.1 Per-Layer RSA and Procrustes Trajectories

Table 12 shows that RSA correlation generally increases with layer depth for all models, with the steepest gains occurring between L3 and L8. BERT-wwm shows a steeper trajectory than baseline BERT, confirming that WWM training accelerates the emergence of cross-scale geometric alignment. RSA peaks around the upper layers, with L11 slightly above the L12 layer used for D3 causal interventions.

Table 13 provides a same-layer supplementary check for the Procrustes diagnostic. The Procrustes trajectories are less monotonic than the RSA trajectories, but they support the same broad interpretation: BERT-wwm maintains consistently low disparity in the upper layers, while baseline BERT shows weaker and less stable geometric alignment.

D.2.2 Statistical Power Analysis

For the RSA analysis (D2), the sample size is inherently limited by the number of categories ($N = 7$), which yields only $7\times6/2 = 21$ unique pairwise distances in each Representational Dissimilarity Matrix. With only 21 distances, the RSA test is underpowered for medium effects and becomes reliable mainly for large correlations. BERT's non-significant RSA result ($\rho = 0.377$, $p = 0.114$) should therefore be interpreted cautiously rather than as a clean null, especially because the corresponding Procrustes result is marginal ($p = 0.056$). By contrast, BERT-wwm's RSA effect is large enough ($\rho = 0.896$) that statistical power is not a serious concern.

Table 10: Per-category recall for D1 cross-scale transfer at L12. Values are means over five random seeds.

Model	Direction	subj.-pred.	prep.-obj.	mod.-head	adv.-head	serial	verb-obj.	verb-comp.
BERT	w→s	9.8	22.8	71.2	19.1	30.9	22.3	35.2
BERT	s→w	31.7	31.3	60.0	14.0	47.5	9.4	7.9
BERT-wwm	w→s	17.6	48.4	61.7	13.5	25.8	34.4	30.2
BERT-wwm	s→w	12.1	16.6	70.2	10.2	69.8	12.8	2.3
ERNIE	w→s	3.3	10.6	50.0	31.6	11.2	45.8	23.1
ERNIE	s→w	6.4	21.1	65.7	16.6	64.9	19.6	7.9
RoBERTa	w→s	15.1	23.4	45.9	2.9	22.6	12.1	52.3
RoBERTa	s→w	15.8	22.6	58.5	7.2	59.2	4.9	7.5

Table 11: D1 robustness with L12 mean-pooled token representations. Values are five-seed means.

Model	w→s	s→w	w→w	s→s
BERT	27.1	29.8	76.7	80.0
BERT-wwm	32.4	30.8	74.4	80.6
ERNIE	31.6	30.0	75.3	81.0
RoBERTa	30.8	31.3	77.8	79.5

Table 12: Spearman rank correlation (ρ) between word-level and syntax-level RDMs across all layers.

Layer	BERT	BERT-wwm	ERNIE	RoBERTa
L1	0.12	0.18	0.21	0.15
L2	0.15	0.25	0.28	0.22
L3	0.19	0.38	0.35	0.31
L4	0.24	0.52	0.42	0.45
L5	0.28	0.65	0.51	0.58
L6	0.31	0.74	0.56	0.65
L7	0.34	0.81	0.60	0.70
L8	0.36	0.85	0.63	0.73
L9	0.37	0.88	0.64	0.74
L10	0.38	0.89	0.65	0.75
L11	**0.39**	**0.90**	**0.66**	**0.76**
L12	0.377	0.896	0.661	0.755

Table 13: Layer-wise same-layer Procrustes disparity between word-level and syntax-level category prototypes. Lower disparity indicates stronger alignment; permutation p-values are shown in parentheses.

Layer	BERT	BERT-wwm	ERNIE	RoBERTa
L1	0.335 (0.066)	0.297 (0.036)	0.313 (0.064)	0.306 (0.037)
L2	0.218 (0.017)	0.282 (0.071)	0.174 (0.005)	0.197 (0.008)
L3	0.132 (0.002)	0.190 (0.011)	0.206 (0.006)	0.173 (0.007)
L4	0.275 (0.039)	0.163 (0.015)	0.198 (0.006)	0.108 (0.004)
L5	0.183 (0.012)	0.114 (0.002)	0.185 (0.010)	0.128 (0.002)
L6	0.179 (0.010)	0.101 (0.001)	0.173 (0.005)	0.077 (0.000)
L7	0.160 (0.005)	0.101 (0.001)	0.183 (0.006)	0.091 (0.001)
L8	0.174 (0.007)	0.091 (0.001)	0.142 (0.001)	0.091 (0.002)
L9	0.208 (0.016)	0.089 (0.001)	0.229 (0.125)	0.068 (0.000)
L10	0.230 (0.046)	0.092 (0.001)	0.339 (0.318)	0.328 (0.079)
L11	0.234 (0.051)	0.088 (0.001)	0.175 (0.021)	0.118 (0.001)
L12	0.171 (0.048)	0.095 (0.001)	0.177 (0.028)	0.170 (0.092)

E D3: Circuit Isomorphism

E.1 Implementation Details

E.1.1 Gradient-Based Dimension Importance Ranking

For the causal intervention experiment (D3), we identify the dimensions most critical for *the source task of each intervention direction* using a gradient-based importance score. In the w→s direction, the source task is word-level classification; in the s→w direction, it is syntax-level classification. Given the trained task-specific linear probe $f(\cdot)$ used in D1, the importance score for input dimension j is computed as:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \left| \frac{\partial \mathcal{L}(f(h_i), y_i)}{\partial h_{i,j}} \right| \tag{9}$$

where $h_i \in \mathbb{R}^d$ is the representation of the i-th source-task sample, y_i is its category label, $\mathcal{L}$ is the cross-entropy loss, and $h_{i,j}$ is the j-th dimension of h_i. Intuitively, I_j measures how much perturbing input dimension j affects the classification loss on average across all samples and categories.

We then rank all $d = 768$ dimensions by I_j in descending order and select the Top-50 dimensions for intervention. The same 50 dimensions are used for both the *suppress* and *replace* interventions.

E.1.2 Intervention Procedure

We perform two types of intervention on the selected dimensions when processing opposite-scale target inputs: *Suppress*: Scaling down the selected dimension values by a fixed suppression factor; *Replace*: Substituting the selected dimensions with the corresponding source-scale category prototype values.

For w→s, word-critical dimensions and word-category prototypes are applied to syntax-level hidden states and evaluated with the syntax probe. For s→w, syntax-critical dimensions and syntax-category prototypes are applied to word-level hidden states and evaluated with the word probe. To establish a baseline, we perform the same interventions on 50 randomly selected dimensions and repeat this 100 times to obtain a stable random base-

line distribution.

E.1.3 Same-Layer Design

The main D3 analysis uses a same-layer L12 design: source-dimension ranking, target-side intervention, and target-probe evaluation are all performed in the same representational space. This avoids a cross-layer measurement confound in which interventions at L12 would be evaluated by probes trained at model-specific syntax-optimal layers. We report the earlier cross-layer setting separately as a diagnostic comparison.

E.1.4 Affected-Sample Criterion

Statistical significance is determined using Benjamini–Hochberg FDR correction ($q = 0.05$) combined with an effect size threshold ($|\Delta| \geq 0.01$). A sample is counted as significantly affected if both criteria are met.

E.2 Supplementary Results and Robustness Checks

E.2.1 Bidirectional Same-Layer Counts

Table 14 gives the raw affected-sample counts. Both directions show the same pattern: targeted dimensions are disruptive in absolute terms, but not more disruptive than random dimension sets of the same size.

E.2.2 Top-k Sensitivity Analysis

We repeated D3 with smaller source-critical dimension sets ($k = 5, 10, 20$) to check whether the Top-50 intervention made the random baseline too strong. Table 15 shows that targeted interventions remain close to matched random baselines, with fold changes between 0.85 and 1.06.

E.2.3 Suppress vs. Replace Interventions

Table 16 compares suppressing Top-50 word-critical dimensions with replacing them by word-category prototype values. The two intervention types yield nearly identical w→s results.

E.2.4 Cross-Layer Diagnostic Comparison

Table 17 compares the earlier cross-layer setting with the same-layer design used in the main analysis. The conclusion is unchanged: targeted interventions still produce no greater disruption than random baselines.

E.2.5 Attention Head-Level Interventions

We also tested attention-head-level localization by zeroing the Top-10 word-critical heads during syntax-level processing. Table 18 shows the same pattern as the dimension-level results: targeted heads are no more disruptive than random heads.

Table 14: Bidirectional same-layer D3 results with raw affected-sample counts.

Model	w→s			s→w		
	Sig.	Rand.	Fold	Sig.	Rand.	Fold
BERT	832 / 1204 (69.1%)	69.7%	0.99×	265 / 371 (71.4%)	72.4%	0.99×
BERT-wwm	976 / 1204 (81.1%)	78.1%	1.04×	333 / 371 (89.8%)	90.3%	0.99×
ERNIE	1071 / 1204 (89.0%)	85.7%	1.04×	358 / 371 (96.5%)	96.3%	1.00×
RoBERTa	1083 / 1204 (90.0%)	89.9%	1.00×	347 / 371 (93.5%)	92.7%	1.01×

Table 15: Top-k sensitivity for D3. Each Top-k cell reports targeted/random disruption percentages. The fold range summarizes targeted-vs.-random ratios across $k = 5, 10, 20$.

Model	Direction	Top-5	Top-10	Top-20	Fold Range
BERT	w→s	48.7/49.4	55.7/54.9	60.1/61.0	0.99–1.01
BERT	s→w	42.6/50.1	58.2/56.7	61.2/63.2	0.85–1.03
BERT-wwm	w→s	48.0/55.6	61.8/62.8	69.9/68.9	0.86–1.01
BERT-wwm	s→w	63.9/67.4	69.8/77.2	82.2/83.3	0.90–0.99
ERNIE	w→s	63.6/64.0	74.7/72.0	81.7/78.5	0.99–1.04
ERNIE	s→w	52.3/56.8	73.6/75.2	84.9/87.6	0.92–0.98
RoBERTa	w→s	72.3/70.9	79.7/78.8	86.1/83.7	1.01–1.03
RoBERTa	s→w	60.4/57.0	75.7/73.7	87.9/83.6	1.03–1.06

Table 16: Suppress vs. replace intervention-type check for D3 in the w→s direction.

Model	Suppress		Replace		Random Baseline	
	Sig. %	Fold	Sig. %	Fold	Sig. %	Fold
BERT	69.1%	1.0×	68.8%	1.0×	69.7%	1.0×
BERT-wwm	81.1%	1.0×	80.5%	1.0×	78.1%	1.0×
ERNIE	89.0%	1.0×	88.3%	1.0×	85.7%	1.0×
RoBERTa	90.0%	1.0×	89.5%	1.0×	89.9%	1.0×

Table 17: Cross-layer vs. same-layer D3 intervention results.

Model	Cross-layer		Same-layer		Random Baseline	
	Sig. %	Fold	Sig. %	Fold	Sig. %	Fold
BERT	72.3%	1.1×	69.1%	1.0×	69.7%	1.0×
BERT-wwm	84.5%	1.1×	81.1%	1.0×	78.1%	1.0×
ERNIE	91.2%	1.1×	89.0%	1.0×	85.7%	1.0×
RoBERTa	92.1%	1.0×	90.0%	1.0×	89.9%	1.0×

Table 18: Attention head-level D3 intervention results.

Model	Sig. Samples	% of Total	Random Baseline	Fold Increase
BERT	825	68.5%	69.7%	0.98×
BERT-wwm	961	79.8%	78.1%	1.02×
ERNIE	1058	87.9%	85.7%	1.03×
RoBERTa	1071	89.0%	89.9%	0.99×