

Typed Relation Path Geometry in Sentence Embedding Spaces

Mingyue Liu and Ziyang Xing
University of Oxford
Oxford, United Kingdom
{com11079, sjoh5910}@ox.ac.uk

Abstract

Lexical relations are not only labels on pairs of concepts: they are directed typed movements that can form paths. We study whether sentence embedding spaces preserve this structure at the level of individual relation instances. By representing two-hop lexical chains as composed movements on the normalised embedding sphere, we test whether true chains form more coherent routes than matched false-middle alternatives. Across controlled lexical path experiments, real chains exhibit a robust path signal, including for mixed relation compositions, and the composed movements retain information about ordered relation type. Beyond diagnosis, the same path geometry aligns with independent distributional context shifts and improves graded lexical-entailment transfer, indicating that it captures reusable lexical structure rather than a task-specific artefact.

1 Introduction

Sentence embeddings are usually treated as proximity spaces: expressions that lie close together are taken to be semantically related, while distant expressions are taken to be less related. This view has made frozen sentence encoders useful for retrieval, clustering, and semantic search. However, proximity captures only neighbourhood structure. Many semantic phenomena are not merely neighbourhood-based, but directed, typed, and compositional.

Lexical relations provide a clear example. A *kitten* moves to *cat* by a kind-of relation (hypernymy); a *whisker* moves to *cat* by a part-of relation (holonymy). These two edges share an endpoint, but they are not the same kind of semantic movement. Moreover, they can participate in longer chains:

$$\text{whisker} \xrightarrow{\text{part-of}} \text{cat} \xrightarrow{\text{kind-of}} \text{mammal}.$$

Such a chain is not merely a set of related concepts.

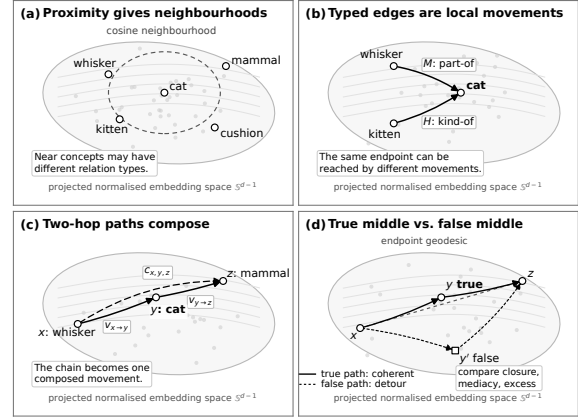


Figure 1: Overview of typed lexical path geometry. Concepts are shown as points in a projected normalised embedding space, relation instances as directed local movements, and two-hop lexical chains as composed movements. The key evaluation compares the route through the true middle y with matched false-middle routes y' , testing whether real typed chains form more coherent paths in embedding space.

It has an ordered typed structure: one relation instance is followed by another. Figure 1 illustrates the central idea of the paper. Proximity can reveal a neighbourhood around *cat*, but it does not by itself determine whether the route through *cat* is geometrically privileged, whether another middle concept would break the path, or whether reversing the relation sequence changes the movement. This paper asks whether frozen sentence embedding spaces preserve traces of such typed lexical paths. This question sits between two established perspectives. Work on sentence embedding geometry has mainly studied neighbourhoods, semantic regions, or interpretable feature directions, while relation-specialised models explicitly train vector spaces to encode hierarchy, order, or knowledge-graph transformations. Our focus is different: we treat typed lexical chains as geometric paths in an off-the-shelf frozen sentence encoder. The diagnos-

tic question is whether individual relation instances already behave like local directed movements, and whether these movements compose along short typed chains.

Our basic object is a two-hop path

$$x \xrightarrow{r_1} y \xrightarrow{r_2} z,$$

where r_1 and r_2 may be the same or different relation types. We represent each edge as a local movement and compose the two movements into a candidate route from x to z . This shifts the analysis from *points* to *paths*: rather than asking only whether x and z are close, or whether a classifier can label $x \rightarrow y$, we ask whether the route through the true middle y is more coherent than matched routes through false middles y' . We measure these complementary aspects using three path-level criteria: *closure*, which tests endpoint reconstruction; *mediacy*, which tests whether the middle lies near the endpoint geodesic; and *excess*, which tests whether the two-hop route is efficient rather than a detour. These are path-level criteria: they do not reduce the problem to nearest-neighbour retrieval or pairwise relation classification. While our primary object is the two-hop path, we also analyse its constituent one-hop relations to verify directionality and relation specificity, and to test transfer to graded lexical entailment.

The experiments are designed to test where this geometry appears, how robust it is, and how far it extends. We evaluate controlled lexical paths, mixed relation compositions, matched false-middle controls, semantic-vs-null comparisons, and ordered composition prediction. We further compare the path signal with distributional context profiles and evaluate transfer to lexical entailment and sentence-level inference, allowing us to identify both the reach and the limits of the observed geometry.

The results support a positive and bounded answer. Across frozen encoders, real typed chains are more coherent than matched false-middle chains under closure, mediacy, and excess, showing that the true middle is geometrically privileged. Null comparisons indicate that the effect depends on meaningful lexical alignment rather than unit-sphere geometry alone. Composed path vectors also predict ordered relation-composition types, suggesting that the signal is not reducible to endpoint relatedness. Additional analyses connect the effect to distributional context shifts and show use-

ful transfer to lexical entailment, while sentence-level inference experiments mark its boundary.

This paper makes three contributions. First, it introduces typed lexical paths as a diagnostic object for analysing frozen sentence embedding spaces beyond point proximity and pairwise relation direction. Second, it develops a geometric audit framework based on composed movements, false-middle controls, null comparisons, and ordered composition prediction. Third, it provides a bounded empirical account of where typed path geometry appears, how it relates to distributional context shifts, and where its explanatory scope ends.

2 Related Work

2.1 Geometry in sentence representation spaces

Sentence embeddings are typically used as proximity spaces, where cosine similarity or nearest-neighbour structure supports semantic search, clustering, and retrieval (Reimers and Gurevych, 2019; Cer et al., 2018; Conneau et al., 2017). Beyond this operational view, prior work has shown that embedding spaces exhibit substantial global structure, including dominant common directions, anisotropy, and distributional distortions that motivate common-component removal, flow-based normalisation, and whitening (Mu and Viswanath, 2018; Li et al., 2020; Su et al., 2021). More recent studies further examine sparse dimensions, semantic subspaces, feature directions, and feature composition in sentence representations (Trifonov et al., 2018; Opitz and Frank, 2022; Tehenan, 2025). Our work builds on this structural view of embedding geometry, but shifts the focus from global axes or regions to local typed movements and their path-level coherence.

2.2 Relation directions and lexical entailment

Relation-sensitive geometry has been widely studied in word representations. Vector-offset analogies suggest that some syntactic and semantic relations can be expressed as directions, although later work showed that such regularities are relation-dependent and sensitive to evaluation protocol (Mikolov et al., 2013; Levy and Goldberg, 2014; Gladkova et al., 2016). Semantic-axis methods similarly use embedding directions to encode interpretable lexical dimensions (An et al., 2018). Lexical entailment and hypernymy provide a stricter setting because they are asymmetric: distributional

inclusion, hierarchical embeddings, and vector-specialisation methods all aim to recover or impose directional entailment structure (Geffet and Dagan, 2005; Nguyen et al., 2017; Vulić and Mrkšić, 2018; Glavaš and Vulić, 2019), while graded benchmarks show that such relations vary continuously in strength (Vulić et al., 2017). These works primarily evaluate pairwise relation detection, directionality, or specialisation; we instead test whether relation movements compose across multi-step lexical paths.

2.3 Pair-level relation representations

A related line of work represents relations directly rather than deriving them from isolated word or sentence embeddings. Relation-specific word embeddings aim to separate relational information from attributional similarity, addressing the tendency of ordinary word embeddings to conflate different forms of lexical association (Camacho-Collados et al., 2019). Word-pair representations similarly encode implicit relations between two words and have been used for cross-sentence inference (Joshi et al., 2019). More recent relation encoders distil pair-level relation embeddings from pretrained models and fine-tune them so that relationally similar pairs are close in the learned space (Ushio et al., 2021, 2025). These models make relations explicit representational objects. In contrast, we do not train a relation encoder or pair-level embedding space; we probe whether sentence encoders already contain relation-like movements that support typed path composition.

2.4 Learned relational and path geometries

Relational geometry is explicit in knowledge representation learning. Tensor, bilinear, translational, and rotation-based models encode relations through learned interactions, translations, or rotations between entity embeddings (Nickel et al., 2011; Trouillon et al., 2016; Bordes et al., 2013; Sun et al., 2019). Hierarchical and order-based models impose stronger geometric biases, using partial orders, hyperbolic spaces, or entailment cones to capture asymmetric and tree-like structure (Vendrov et al., 2016; Nickel and Kiela, 2017; Ganea et al., 2018). Multi-hop relation-path models further learn or compose relation sequences for knowledge-base inference (Lin et al., 2015; Toutanova et al., 2016). Unlike these graph-trained models, our setting is more diagnostic: we ask whether independently trained sentence encoders exhibit geometric traces

of typed lexical path composition without relation-specific training or path-supervised objectives.

3 Typed Path Geometry in Embedding Spaces

This section defines the geometric representation used throughout the paper. Our aim is to formulate typed lexical relation paths as compositional movements. The construction has four steps: normalisation removes embedding magnitude, angular geometry represents relationships on the unit sphere, the logarithm map represents each directed edge as a local movement, and parallel transport places successive movements in a common tangent space so that they can be composed.

3.1 Embeddings on the unit hypersphere

Let $f(x) \in \mathbb{R}^d$ denote the sentence embedding of a concept x . We analyse their length-normalised representations:

$$\hat{f}(x) = \frac{f(x)}{\|f(x)\|_2} \in \mathbb{S}^{d-1}.$$

This places all concepts on the unit hypersphere. The corresponding intrinsic angular distance is

$$d_{\mathbb{S}}(x, y) = \arccos \left(\hat{f}(x)^\top \hat{f}(y) \right).$$

This choice matches the empirical use of sentence embeddings in retrieval, where cosine similarity or dot products between normalised vectors are standard.

3.2 Relation edges as tangent movements

A symmetric similarity score cannot distinguish $x \rightarrow y$ from $y \rightarrow x$. For asymmetric lexical relations, we instead represent an edge as a movement. The simplest movement is the Euclidean displacement

$$\delta_{x \rightarrow y} = \hat{f}(y) - \hat{f}(x). \quad (1)$$

However, if embeddings are analysed on the unit sphere, the corresponding local movement is the Riemannian logarithm map:

$$v_{x \rightarrow y} = \log_{\hat{f}(x)} \hat{f}(y). \quad (2)$$

This vector lies in the tangent space at $\hat{f}(x)$. Its norm is the spherical distance from x to y , and its direction is the local direction of movement

from the source to the target. For two unit vectors $p, q \in \mathbb{S}^{d-1}$, the logarithm map is

$$\log_p(q) = \frac{\theta}{\sin \theta} (q - \cos \theta p), \quad \theta = \arccos(p^\top q). \quad (3)$$

The inverse operation is the exponential map:

$$\exp_p(v) = \cos(\|v\|)p + \sin(\|v\|) \frac{v}{\|v\|}. \quad (4)$$

3.3 Typed two-hop paths

For a typed two-hop path

$$x \xrightarrow{r_1} y \xrightarrow{r_2} z, \quad (5)$$

where r_1 and r_2 correspond to the two relation movements $v_{x \rightarrow y}$ and $v_{y \rightarrow z}$ respectively, the first edge movement $v_{x \rightarrow y}$ lies in the tangent space at $\hat{f}(x)$, whereas the second edge movement $v_{y \rightarrow z}$ lies in the tangent space at $\hat{f}(y)$. These vectors cannot be added directly because they are defined in different local vector spaces. We therefore parallel-transport the second movement to the tangent space at $\hat{f}(x)$:

$$\tilde{v}_{y \rightarrow z} = \text{PT}_{\hat{f}(y) \rightarrow \hat{f}(x)} v_{y \rightarrow z}. \quad (6)$$

For non-antipodal $p, q \in \mathbb{S}^{d-1}$ and $v \in T_p \mathbb{S}^{d-1}$, we use parallel transport along the shortest geodesic from p to q :

$$\text{PT}_{p \rightarrow q}(v) = v - \frac{v^\top q}{1 + p^\top q} (p + q). \quad (7)$$

In Equation (6), $p = \hat{f}(y)$ and $q = \hat{f}(x)$, so the second edge movement is expressed in the tangent space at $\hat{f}(x)$ before it is added to the first movement.

The composed path vector is then

$$c_{x,y,z} = v_{x \rightarrow y} + \tilde{v}_{y \rightarrow z}. \quad (8)$$

Equation 8 is spherical analogue of adding two relation movements. If the embedding space preserves typed relation paths, then moving from x to y and then from y to z should approximately reproduce the direct semantic movement from x to z .

3.4 Path coherence: closure, mediacy, and excess

The composed vector $c_{x,y,z}$ defines a candidate endpoint reached by following the composed movement from x :

$$\hat{z}_{\text{comp}} = \exp_{\hat{f}(x)}(c_{x,y,z}).$$

We evaluate path coherence using three complementary quantities.

Closure. Closure measures the deviation of the recovered endpoint with composed movement from the true endpoint:

$$\varepsilon_C(x, y, z) = d_{\mathbb{S}}(\hat{z}_{\text{comp}}, \hat{f}(z)).$$

Low closure error means that the two edge movements compose into an endpoint close to z .

Mediacy. Mediacy measures the deviation of the middle point from the shortest spherical path between the endpoints. Let $\gamma_{x,z}(t)$, $t \in [0, 1]$, denote the geodesic from $\hat{f}(x)$ to $\hat{f}(z)$:

$$\varepsilon_M(x, y, z) = \min_{t \in [0,1]} d_{\mathbb{S}}(\hat{f}(y), \gamma_{x,z}(t)).$$

Low mediacy error means that y lies close to the endpoint trajectory from x to z . Closure alone is insufficient, because a composed movement may reach the endpoint accidentally even if the middle concept is not geometrically meaningful.

Excess. Excess measures the excess distance covered by the route through y from the geodesic path:

$$\varepsilon_E(x, y, z) = d_{\mathbb{S}}(x, y) + d_{\mathbb{S}}(y, z) - d_{\mathbb{S}}(x, z).$$

Low excess means that the two-hop route through y is close to the direct route from x to z .

These metrics provide complementary, but not independent, views of path coherence. Closure evaluates endpoint reconstruction, mediacy evaluates the placement of the middle concept relative to the endpoint geodesic, and excess evaluates route efficiency. Because all three are functions of the same spherical triangle, they should be interpreted as three jointly satisfied path properties rather than as independent statistical confirmations.

To evaluate how meaningful the real middle y is, we define a false middle y' as a replacement intermediate concept selected under controlled conditions. Specifically, we compare a real path $x \rightarrow y \rightarrow z$ with a false-middle path $x \rightarrow y' \rightarrow z$, where the endpoints x and z are fixed but the middle concept is deliberately substituted.

We then define metrics to evaluate the performance of the path via the true middle compared to paths via false middles.

Advantages. For a false middle y' , we define the advantage of the true path as the reduction in error obtained by using the true middle y rather than y' :

$$\Delta_C = \varepsilon_C(x, y', z) - \varepsilon_C(x, y, z), \quad (9)$$

$$\Delta_M = \varepsilon_M(x, y', z) - \varepsilon_M(x, y, z), \quad (10)$$

$$\Delta_E = \varepsilon_E(x, y', z) - \varepsilon_E(x, y, z). \quad (11)$$

Compared to the false middle, a positive closure advantage means that the true middle gives a composed path that better reaches the endpoint, a positive mediacy advantage means that the true middle lies closer to the endpoint geodesic, and a positive excess advantage means that the true middle gives a shorter and more efficient two-hop route.

4 Experimental Setup

The previous section defined typed lexical paths, composed tangent movements, and the three path-coherence advantages used throughout the paper. We now describe how this framework is instantiated empirically. The experimental design has three roles. First, it tests the main claim under controlled lexical paths: real typed chains should be more coherent than matched false-middle alternatives. Second, it checks whether the effect depends on the semantic organisation of embeddings rather than on generic properties of normalised vectors. Third, it examines whether the resulting geometry is reusable beyond the diagnostic path task, through composition-type prediction, distributional profile probes, and lexical transfer.

Research questions. The experiments are organised as a sequence of increasingly specific tests aim to answer the following research questions:

- Q1:** Do real typed WordNet paths show positive closure, mediacy, and excess advantages under the strongest false-middle control?
- Q2:** Does the pattern survive outside WordNet on a sparse commonsense-relation resource?
- Q3:** Is the effect semantic, or can it be produced by generic unit-sphere geometry?
- Q4:** Do composed path vectors retain information about the ordered relation-composition type?
- Q5:** How far does the geometry transfer, and where does it stop?

4.1 Resources

Table 1 summarises the resources and their roles. H denotes hypernymy and M denotes meronymy. A path label $R_1 \Rightarrow R_2$ denotes an ordered relation-type composition: the first relation movement $x \rightarrow y$ has type R_1 , and the second relation movement $y \rightarrow z$ has type R_2 . We use $\Rightarrow$, rather than $\rightarrow$, to distinguish ordered relation-type compositions from paths between individual concepts.

WordNet (Miller, 1995) is the main benchmark because it provides typed lexical relations and balanced two-hop path families. For noun synsets, hy-

pernym edges are oriented from specific to general. The M relation is oriented from part to whole and is constructed from WordNet’s part, member, and substance holonym links. We form a typed two-hop path $x \xrightarrow{r_1} y \xrightarrow{r_2} z$ whenever the directed typed edges (x, r_1, y) and (y, r_2, z) are both present.

For ConceptNet (Speer et al., 2017), we retain English assertions for AtLocation, CapableOf, and Causes from the static assertions file, normalise endpoint labels, and remove self-loops and duplicate edges. Paths are then formed by the same directed-edge join. Because typed-path coverage is uneven after filtering, ConceptNet is used as an external robustness check rather than as a balanced benchmark, and we report only path families with sufficient coverage.

Brown-corpus (Francis and Kučera, 1964) term frequency–inverse document frequency (TF–IDF) context profiles provide an independent distributional space for testing whether relation paths resemble additive profile shifts. HyperLex (Vulić et al., 2017) tests lexical transfer, while MNLI and HANS (Williams et al., 2018; McCoy et al., 2019) test whether the lexical geometry should or should not be interpreted as sentence-level entailment geometry.

4.2 Models and representations

We evaluate four sentence encoders: all-MiniLM-L6-v2, all-mpnet-base-v2, E5-base-v2, and BGE-base-en-v1.5 (Wang et al., 2020; Song et al., 2020; Wang et al., 2022; Xiao et al., 2024). The encoders are not fine-tuned. This is essential to the paper’s purpose: we audit geometric structure already present in off-the-shelf sentence embeddings rather than training a task-specific model.

The main representation mode is lemma-neutral. For each concept lemma w , we construct five relation-neutral inputs:

- $\{w\}$.
- The term is $\{w\}$.
- They mentioned $\{w\}$.
- The word $\{w\}$ appears.
- People discussed $\{w\}$.

We encode the five inputs independently, average the resulting vectors, and then ℓ_2 -normalise the average. The templates avoid relation-revealing words such as *kind*, *category*, *part*, and *whole*.

Gloss-mode embeddings are reported only as an appendix ablation because glosses can introduce definitional leakage.

Resource	Item	Count	Role in the experiments
WordNet	hypernym edges	5,000	main typed lexical relation
WordNet	meronym edges	3,800	main typed lexical relation
WordNet	$H \Rightarrow H$ paths	1,100	same-relation path family
WordNet	$H \Rightarrow M$ paths	1,100	mixed relation path family
WordNet	$M \Rightarrow H$ paths	1,100	mixed relation path family
WordNet	$M \Rightarrow M$ paths	1,100	same-relation path family
ConceptNet	AtLocation edges	1,500	external commonsense relation
ConceptNet	CapableOf edges	1,500	external commonsense relation
ConceptNet	Causes edges	1,500	external commonsense relation
ConceptNet	AtLocation $\Rightarrow$ AtLocation paths	542	supported ConceptNet path family
ConceptNet	AtLocation $\Rightarrow$ CapableOf paths	162	supported ConceptNet path family
ConceptNet	all other paths	72	too small for meaningful evaluation
Brown context profiles	covered entities	5,358	independent distributional profile space
Brown context profiles	profile-covered WordNet paths	506	profile-space explanation
HyperLex	graded lexical-entailment pairs	2,616	external lexical transfer
MNLI	validation sentence pairs	1,500	sentence-level boundary test
HANS	sentence pairs	1,400	sentence-level boundary test

Table 1: Resources used in the experiments. Hypernymy (H) is oriented from specific to general, and meronymy (M) from part to whole. MNLI denotes Multi-Genre Natural Language Inference; HANS denotes Heuristic Analysis for NLI Systems.

4.3 False-middle controls and reporting

For every resource-attested path $x \xrightarrow{r_1} y \xrightarrow{r_2} z$, the endpoints x and z remain fixed, and the resource-selected middle y is replaced by a distinct candidate y' . Candidate middles are drawn from middle concepts occurring in other paths in the relevant evaluation pool.

For distance-based matching, we define the discrepancy of a candidate as

$$D(y') = |d_S(x, y') - d_S(x, y)| + |d_S(y', z) - d_S(y, z)|. \quad (12)$$

We construct five controls:

Shuffled. Randomly reassign a middle concept from another path in the evaluation pool.

Depth/lexname-matched. For WordNet, sample a candidate with the same lexname as y and with $|h(y') - h(y)| \leq 1$, where $h(\cdot)$ denotes synset depth.

Distance-matched. Select the candidate that minimises $D(y')$ in Equation 12.

Relation-type-matched. Sample a middle from another path with the same ordered relation-composition family (r_1, r_2) .

Relation-type-distance-matched. Within the relation-type-matched candidate pool, select the candidate that minimises $D(y')$.

The main analyses use the relation-type-distance-matched control, the most restrictive of the five controls. It draws the replacement middle from the same ordered relation-composition family

and minimises deviation from the original path’s endpoint-distance profile.

Relation-type matching constrains the pool from which y' is drawn; graph membership with the fixed endpoints is not part of the original matching rule. Appendix A.1 reports a separate robustness analysis that identifies and replaces candidates that themselves complete an attested path with the same endpoints and ordered relation types.

All path advantages are reported in degrees. Positive values mean that the resource-selected middle is more coherent than the matched replacement middle.

5 Results and Analysis

5.1 A1: WordNet paths are coherent under the hardest control

A1 evaluates the central path-coherence claim in the balanced WordNet setting. Table 2 gives the central result. All four path families are positive for all four encoders and all three metrics, even though the false middle is matched by both relation-composition type and endpoint-distance profile. This directly supports A1: the true middle contributes path information beyond pairwise proximity and broad relation-family membership.

The strongest effects are not limited to same-relation chains. Mixed paths are particularly informative because they require two relation types to interact. $H \Rightarrow M$ is consistently stronger than $M \Rightarrow H$, especially for MiniLM, MPNet, and

Path	Model	n	Δ_C°	Δ_M°	Δ_E°
$H \Rightarrow H$	MiniLM	1100	2.34 $\pm$ 0.21	6.11 $\pm$ 0.53	7.95 $\pm$ 0.72
$H \Rightarrow H$	MPNet	1100	2.17 $\pm$ 0.17	6.03 $\pm$ 0.48	8.25 $\pm$ 0.66
$H \Rightarrow H$	E5-base	1100	0.20 $\pm$ 0.02	2.38 $\pm$ 0.19	3.61 $\pm$ 0.29
$H \Rightarrow H$	BGE-base	1100	0.98 $\pm$ 0.08	4.23 $\pm$ 0.31	6.12 $\pm$ 0.47
$H \Rightarrow M$	MiniLM	1100	3.61 $\pm$ 0.24	9.22 $\pm$ 0.64	11.33 $\pm$ 0.86
$H \Rightarrow M$	MPNet	1100	2.88 $\pm$ 0.21	7.84 $\pm$ 0.66	9.85 $\pm$ 0.86
$H \Rightarrow M$	E5-base	1100	0.37 $\pm$ 0.02	4.00 $\pm$ 0.27	5.56 $\pm$ 0.39
$H \Rightarrow M$	BGE-base	1100	1.91 $\pm$ 0.11	7.51 $\pm$ 0.46	9.86 $\pm$ 0.62
$M \Rightarrow H$	MiniLM	1100	2.31 $\pm$ 0.26	5.82 $\pm$ 0.67	7.04 $\pm$ 0.82
$M \Rightarrow H$	MPNet	1100	1.96 $\pm$ 0.23	5.27 $\pm$ 0.61	6.63 $\pm$ 0.76
$M \Rightarrow H$	E5-base	1100	0.25 $\pm$ 0.03	2.76 $\pm$ 0.29	3.72 $\pm$ 0.37
$M \Rightarrow H$	BGE-base	1100	1.25 $\pm$ 0.13	4.92 $\pm$ 0.51	6.21 $\pm$ 0.62
$M \Rightarrow M$	MiniLM	1100	2.09 $\pm$ 0.21	5.50 $\pm$ 0.55	6.73 $\pm$ 0.66
$M \Rightarrow M$	MPNet	1100	1.67 $\pm$ 0.21	4.69 $\pm$ 0.52	5.96 $\pm$ 0.60
$M \Rightarrow M$	E5-base	1100	0.25 $\pm$ 0.03	3.03 $\pm$ 0.27	4.21 $\pm$ 0.35
$M \Rightarrow M$	BGE-base	1100	1.32 $\pm$ 0.13	5.48 $\pm$ 0.50	7.28 $\pm$ 0.60

Table 2: WordNet path coherence under relation-type-distance-matched controls. Values are angular advantages in degrees, reported as mean $\pm$ half-width of the 95% confidence interval. Positive values mean that the true middle gives a more coherent path than the matched false middle. Bold marks the strongest value within each path block and metric. Here, n is the number of evaluated path instances for each encoder and path family.

BGE-base. This asymmetry matters: if the metrics were only detecting generic relatedness, the ordered composition should not matter as much. The result therefore supports a typed and ordered path interpretation, while still showing model-dependent magnitude. E5-base preserves the same qualitative pattern but with smaller closure advantages.

5.2 A2: ConceptNet provides external support

A2 tests whether this pattern is confined to WordNet. We apply the same diagnostic to the retained ConceptNet path groups. Table 3 shows the same positive pattern for the two path families with sufficient coverage.

5.3 A3: Null embeddings remove the path signal

A3 controls for geometry without semantics. We test whether the observed advantages require meaningful concept–vector alignment. Table 4 addresses this by comparing semantic embeddings with two nulls: a permuted null, which keeps the real vectors but randomly reassigns them to concepts, and a random-unit null, which assigns each concept to an independent random point on the unit sphere. We then compare the results of the same concepts under these embeddings.

The null effects are near zero, whereas the semantic embeddings retain multi-degree mediacy and excess advantages. This answers Q3: path

coherence depends on the alignment between concepts and their semantic vectors, not merely on the geometry of normalised random vectors.

5.4 A4: Composed vectors retain relation-composition type

Q4 asks whether the composed vector $c_{x,y,z}$ also carries information about the ordered composition type. We therefore train a shallow classifier to distinguish $H \Rightarrow H$, $H \Rightarrow M$, $M \Rightarrow H$, and $M \Rightarrow M$. We also evaluate whether this survives stricter disjoint and held-out splits.

Table 5 shows above-chance macro-F1 across all splits. The splits consist of standard split, single concept disjoint, entity disjoint, and held-outs. A disjoint evaluation means none of the specified concepts overlap both training and testing sets, while a standard split has no such limitations. For example, a y -disjoint split indicates that the training and testing sample have no overlapping y (the middle concept of $x \rightarrow y \rightarrow z$), while entity-disjoint split has no overlap for any x , y , and z combination. The lexname-held-out evaluation excludes entire lexical domains from training, and reserves them for testing. Performance remains well above 0.25 under all splits, indicating that the signal is not only memorisation of a single concept position. While still outperforming the baseline, the lexname-held-out split is substantially harder. The composed vectors therefore contain typed-composition information, but this information is still domain-sensitive rather

Path	Model	n	Δ_C°	Δ_M°	Δ_E°
AtLoc. $\Rightarrow$ AtLoc.	MiniLM	542	3.03 $\pm$ 0.32	5.66 $\pm$ 0.62	7.84 $\pm$ 0.79
AtLoc. $\Rightarrow$ AtLoc.	MPNet	542	2.50 $\pm$ 0.24	4.95 $\pm$ 0.53	7.05 $\pm$ 0.66
AtLoc. $\Rightarrow$ AtLoc.	E5-base	542	0.26 $\pm$ 0.03	2.42 $\pm$ 0.25	3.56 $\pm$ 0.33
AtLoc. $\Rightarrow$ AtLoc.	BGE-base	542	1.21 $\pm$ 0.11	3.99 $\pm$ 0.39	5.66 $\pm$ 0.52
AtLoc. $\Rightarrow$ CapOf.	MiniLM	162	4.71 $\pm$ 0.58	8.35 $\pm$ 1.05	11.96 $\pm$ 1.46
AtLoc. $\Rightarrow$ CapOf.	MPNet	162	5.20 $\pm$ 0.61	9.76 $\pm$ 1.32	14.26 $\pm$ 1.96
AtLoc. $\Rightarrow$ CapOf.	E5-base	162	0.42 $\pm$ 0.05	3.59 $\pm$ 0.51	5.65 $\pm$ 0.75
AtLoc. $\Rightarrow$ CapOf.	BGE-base	162	1.86 $\pm$ 0.21	5.46 $\pm$ 0.85	8.11 $\pm$ 1.04

Table 3: ConceptNet path coherence under relation-type-distance-matched controls. Values are angular advantages in degrees, reported as mean $\pm$ half-width of the 95% confidence interval. AtLoc. abbreviates AtLocation; CapOf. abbreviates CapableOf. Here, n is the number of evaluated path instances for each encoder and path family.

Source	Embedding	Δ_C°	Δ_M°	Δ_E°
WordNet	semantic	1.59	5.29	6.89
WordNet	permuted null	0.01	0.03	0.03
WordNet	random-unit null	0.00	0.00	0.00
ConceptNet	semantic	2.04	4.79	6.84
ConceptNet	permuted null	0.10	0.12	0.19
ConceptNet	random-unit null	0.03	0.02	0.03

Table 4: Semantic versus null embeddings under relation-type-distance controls. Values are means over the four encoders; full per-encoder values are reported in the appendix. Null magnitudes are close to zero compared with semantic embeddings.

Split	MiniLM	MPNet	E5	BGE
Random	0.69	0.75	0.66	0.70
x -disj.	0.68	0.75	0.65	0.70
y -disj.	0.63	0.70	0.61	0.66
z -disj.	0.51	0.62	0.54	0.53
Entity-disj.	0.52	0.61	0.53	0.51
Lexname-held	0.36	0.40	0.39	0.41

Table 5: WordNet relation-composition type prediction from composed path vectors. Values are macro-F1; chance macro-F1 is 0.25. Bold marks the best encoder in each split. E5 abbreviates E5-base, BGE abbreviates BGE-base, disj. abbreviates disjoint, and held abbreviates held-out.

than fully symbolic or domain-invariant.

5.5 A5a: Brown profiles support a contextual-shift explanation

A5a examines one possible source of the observed geometry. If P_x , P_y , and P_z are distributional profiles, then profile differences compose algebraically: $(P_y - P_x) + (P_z - P_y) = P_z - P_x$. We test whether typed paths also appear as coherent shifts in an independent Brown-corpus TF-IDF context-profile space.

Table 6 shows positive mediacy and excess advantages in Brown profile space. This does not

Path	n	Δ_M°	Δ_E°
$H \Rightarrow H$	242	2.69 $\pm$ 1.01	3.96 $\pm$ 1.49
$H \Rightarrow M$	127	8.05 $\pm$ 2.64	11.53 $\pm$ 3.07
$M \Rightarrow H$	63	11.47 $\pm$ 7.37	13.74 $\pm$ 7.85
$M \Rightarrow M$	74	13.54 $\pm$ 4.19	17.72 $\pm$ 4.78

Table 6: Brown-corpus context-profile path advantages under relation-type profile controls. Values are angular advantages in degrees, reported as mean $\pm$ half-width of the 95% confidence interval. Brown profiles are used as an independent distributional probe, not as a competing sentence encoder. Here, n is the number of evaluated path instances for each encoder and path family.

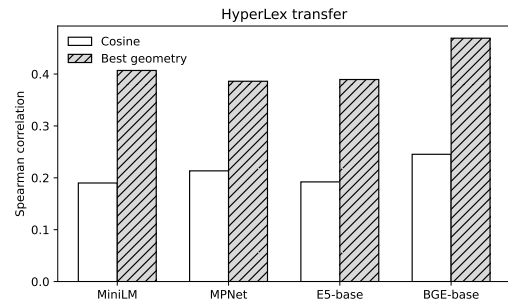


Figure 2: HyperLex transfer. Each pair compares cosine similarity with the best WordNet-derived geometric score for that encoder.

prove that the encoders explicitly store Brown-style profiles. It does, however, support a distributional interpretation: typed relation paths may be coherent because they track additive shifts in contextual evidence.

5.6 A5b: Geometry transfers lexically but not to sentence-level entailment

A5b tests the boundary of the geometry. We ask whether WordNet-derived geometry transfers to an external lexical task. As shown in Figure 2, the best geometric score improves over cosine similarity for all four encoders: MiniLM improves from

Dataset	Cosine AUC	Field AUC	Asymmetry AUC
MNLI	0.78	0.53	0.53
HANS	0.56	0.47	0.48

Table 7: Sentence-level entailment boundary. Values are area under the receiver operating characteristic curve (AUC), averaged across four encoders. Entailment fields trained on MNLI do not transfer robustly to HANS.

0.19 to 0.41, MPNet from 0.21 to 0.39, E5-base from 0.19 to 0.39, and BGE-base from 0.25 to 0.47 in Spearman correlation. This shows that the diagnostic geometry carries useful lexical-entailment information beyond cosine similarity.

The transfer boundary is equally important. Table 7 shows that a global sentence-level entailment field trained on MNLI does not transfer robustly to HANS. Cosine similarity is stronger on MNLI and still weak on HANS, while the learned field and asymmetry scores are near or below chance on HANS. These results support partial typed lexical-path structure, but not a general symbolic calculus for sentence-level inference.

6 Discussion and Conclusion

This paper studied a simple question that is largely missed by proximity-based analyses of sentence embeddings: whether typed lexical relations behave not only as pairwise associations, but as local movements that can be followed. The results give a bounded positive answer. In controlled lexical paths, the real middle concept consistently yields a more coherent route than matched alternatives, even when the false middle is chosen to preserve relation family and endpoint-distance profile. This suggests that the middle of a lexical chain is not interchangeable with any nearby or relation-compatible concept; it occupies a position that is geometrically informative for the ordered path.

The main implication is that sentence embedding spaces can preserve a weak form of relational compositionality without being trained for it. Prior work has shown that embeddings contain neighbourhoods, semantic axes, relation directions, or learned relational transformations. Our findings add a different level of structure: individual typed edges can leave local geometric traces, and pairs of such traces can sometimes compose into a measurable route. The ordered-composition results further indicate that this signal is not exhausted by endpoint similarity, since composed movements retain

information about the type and order of the relation sequence.

The three- and four-hop analysis in Appendix A.10 provides initial evidence that the pattern extends beyond two-hop paths. Future work should test this more systematically across longer and more varied chains, broader relation inventories, and alternative representations of relation composition.

Limitations

This study is deliberately scoped as a controlled diagnostic of typed lexical path geometry in sentence embeddings. The main experiments use short two-hop paths because they provide a clean setting in which path composition can be evaluated without conflating many sources of error. WordNet offers well-defined hypernymy and meronymy relations for this purpose, while ConceptNet provides an external commonsense check with a broader but sparser relation inventory. This design supports precise conclusions about the tested relation families, but leaves substantially longer chains, systematic length scaling, denser commonsense graphs, multilingual paths, and larger relation inventories as natural extensions.

Acknowledgments

Use of AI assistants. We used AI assistants for language polishing, LaTeX formatting, table reformatting, and revision of explanatory text. The authors determined and verified all research ideas, experiments, code, analyses, numerical results, citations, claims, and final manuscript content.

References

- Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. [SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461. Association for Computational Linguistics.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *Advances in Neural Information Processing Systems*, volume 26, pages 2787–2795. Curran Associates, Inc.
- Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019. [Relational word embeddings](#). In *Proceedings of the 57th Annual Meeting of*

- the Association for Computational Linguistics, pages 3286–3296. Association for Computational Linguistics.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder for English](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. [Supervised learning of universal sentence representations from natural language inference data](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.
- W. Nelson Francis and Henry Kučera. 1964. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University, Providence, RI.
- Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. [Hyperbolic entailment cones for learning hierarchical embeddings](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1646–1655. PMLR.
- Maayan Geffet and Ido Dagan. 2005. [The distributional inclusion hypotheses and lexical entailment](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 107–114. Association for Computational Linguistics.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. [Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn’t](#). In *Proceedings of the NAACL Student Research Workshop*, pages 8–15. Association for Computational Linguistics.
- Goran Glavaš and Ivan Vulić. 2019. [Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. [pair2vec: Compositional word-pair embeddings for cross-sentence inference](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3597–3608. Association for Computational Linguistics.
- Omer Levy and Yoav Goldberg. 2014. [Linguistic regularities in sparse and explicit word representations](#). In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180. Association for Computational Linguistics.
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. [On the sentence embeddings from pre-trained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130. Association for Computational Linguistics.
- Xin Li and Dan Roth. 2002. [Learning question classifiers](#). In *COLING 2002: The 19th International Conference on Computational Linguistics*.
- Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015. [Modeling relation paths for representation learning of knowledge bases](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 705–714. Association for Computational Linguistics.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.
- Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. [Linguistic regularities in continuous space word representations](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Association for Computational Linguistics.
- George A. Miller. 1995. [WordNet: A lexical database for English](#). *Communications of the ACM*, 38(11):39–41.
- Jiaqi Mu and Pramod Viswanath. 2018. [All-but-the-Top: Simple and effective postprocessing for word representations](#). In *International Conference on Learning Representations*.
- Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. [Hierarchical embeddings for hypernymy detection and directionality](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243. Association for Computational Linguistics.
- Maximilian Nickel and Douwe Kiela. 2017. [Poincaré embeddings for learning hierarchical representations](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 6338–6347. Curran Associates, Inc.
- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. [A three-way model for collective learning on multi-relational data](#). In *Proceedings of the 28th International Conference on Machine Learning*, pages 809–816, Bellevue, Washington, USA. Omnipress.

- Juri Opitz and Anette Frank. 2022. [SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). 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 3982–3992. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [MPNet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [ConceptNet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. [Whitening sentence representations for better semantics and faster retrieval](#). *arXiv preprint arXiv:2103.15316*.
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. [RotatE: Knowledge graph embedding by relational rotation in complex space](#). In *International Conference on Learning Representations*.
- Matthieu Tehenan. 2025. [Semantic geometry of sentence embeddings](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11993–12004, Suzhou, China. Association for Computational Linguistics.
- Kristina Toutanova, Xi Victoria Lin, Wen-tau Yih, Hoifung Poon, and Chris Quirk. 2016. [Compositional learning of embeddings for relation paths in knowledge base and text](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1434–1444. Association for Computational Linguistics.
- Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, and Thomas Hofmann. 2018. [Learning and evaluating sparse interpretable sentence embeddings](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 200–210. Association for Computational Linguistics.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. [Complex embeddings for simple link prediction](#). In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR.
- Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. [Distilling relation embeddings from pretrained language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2025. [RelBERT: Embedding relations with language models](#). *Artificial Intelligence*, 347:104359.
- Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. [Order-embeddings of images and language](#). In *International Conference on Learning Representations*.
- Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. [HyperLex: A large-scale evaluation of graded lexical entailment](#). *Computational Linguistics*, 43(4):781–835.
- Ivan Vulić and Nikola Mrkšić. 2018. [Specialising word vectors for lexical entailment](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack: Packed resources for general Chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

A Additional Results

This appendix reports supporting and diagnostic results omitted from the main text for readability. The main paper foregrounds the strongest relation-type-distance control and the external transfer and boundary analyses. Here we include full false-middle controls, edge-level directionality and relation-specificity diagnostics, complete HyperLex transfer results, hard-control binary classification, sentence-level entailment details, ConceptNet coverage and null comparisons, gloss-mode ablations, and Euclidean–Riemannian comparisons.

A.1 Full false-middle controls

The path-geometry experiments compare a real path $x \rightarrow y \rightarrow z$ with a false-middle path $x \rightarrow y' \rightarrow z$. The controls are ordered from weak to strong. *Shuffled* middles test against a random alternative from the evaluation pool. *Depth/lexname* matching controls for broad WordNet position and lexical domain where such metadata are meaningful. *Distance* matching controls for pairwise proximity to the endpoints. *Type* matching draws the false middle from the same ordered relation-composition family. *Type+distance* matching combines the two strongest constraints and is the main control used in the paper.

Tables 8–10 report the full five-control results for WordNet, ConceptNet, and the Text REtrieval Conference (TREC) question-classification hierarchy (Li and Roth, 2002). WordNet is the main typed lexical benchmark. ConceptNet is an external commonsense-relation check. TREC is included only as a simple hierarchy sanity check: its paths connect questions to fine labels and then to coarse labels, so it tests whether the same path-coherence metrics detect a known label hierarchy outside WordNet. TREC is not used as evidence for the paper’s main claim about typed lexical relation composition.

Because only WordNet categorises into lexnames, the depth/lexname control only applies to WordNet.

Alternative-valid-middle audit. A candidate selected by relation-type–distance matching could, in principle, also be an attested middle between the same endpoints. We therefore conducted an additional graph-membership audit over all relation-type–distance-matched WordNet controls. Let E denote the directed typed edge set. We classify a

Model	Control	n	Δ_C°	Δ_M°	Δ_E°	Euc. Δ_M°
BGE-base	Depth/lex.	4400	4.46	14.58	21.12	12.44
BGE-base	Distance	4400	1.84	7.08	9.55	6.20
BGE-base	Type+dist.	4400	1.36	5.53	7.37	4.88
BGE-base	Type	4400	5.16	16.37	23.98	13.91
BGE-base	Shuffled	4400	4.96	15.87	23.21	13.49
E5-base	Depth/lex.	4400	0.82	7.84	11.98	7.45
E5-base	Distance	4400	0.36	3.89	5.51	3.72
E5-base	Type+dist.	4400	0.27	3.04	4.27	2.92
E5-base	Type	4400	1.00	9.21	14.30	8.74
E5-base	Shuffled	4400	0.97	9.06	14.07	8.59
MPNet	Depth/lex.	4400	8.60	18.30	25.57	14.31
MPNet	Distance	4400	2.99	7.80	10.13	6.39
MPNet	Type+dist.	4400	2.16	5.94	7.66	4.91
MPNet	Type	4400	11.94	23.44	33.64	18.16
MPNet	Shuffled	4400	11.78	23.18	33.25	17.93
MiniLM	Depth/lex.	4400	9.51	18.82	25.73	14.49
MiniLM	Distance	4400	3.38	8.37	10.56	6.73
MiniLM	Type+dist.	4400	2.58	6.65	8.25	5.41
MiniLM	Type	4400	12.14	22.50	31.51	17.22
MiniLM	Shuffled	4400	12.01	22.30	31.25	17.05

Table 8: Full WordNet false-middle controls. Values are path-level aggregate angular advantages in degrees. The controls are shuffled, depth/lexname, distance-matched, relation-type-matched, and relation-type-distance-matched.

candidate y' as an alternative valid middle when both $(x, r_1, y') \in E$ and $(y', r_2, z) \in E$.

The audit identified 36 of 17,600 encoder–path control assignments (0.20%) as alternative-valid: 5 for MiniLM, 10 for MPNet, 14 for E5-base, and 7 for BGE-base. For each affected assignment, we retained the original evaluated path, removed the graph-valid candidate, and selected the next-best eligible candidate from the same seeded pool using the original relation-type–distance objective. All unaffected controls were retained.

The aggregate mean advantages changed from 1.60° to 1.60° for closure, from 5.30° to 5.30° for mediacy, and from 6.89° to 6.90° for excess. All 16 encoder–path-family combinations remained positive. The minimum clean advantages, obtained for $H \Rightarrow H$ with E5-base, were 0.20° , 2.38° , and 3.61° , respectively. Thus, excluding alternative valid middles does not change the qualitative conclusion of the original analysis.

A.2 Directionality: positive versus reversed edges

The main experiments analyse two-hop path composition. As a supporting diagnostic, we first test whether individual WordNet relation edges contain directional information. For each relation instance, the positive edge is compared with the reversed edge. Since cosine similarity is symmetric, it can-

Model	Control	n	Δ_C°	Δ_M°	Δ_E°	Euc. Δ_M°
BGE-base	Depth/lex.	776	N/A	N/A	N/A	N/A
BGE-base	Distance	776	1.42	4.84	6.70	4.19
BGE-base	Type+dist.	776	1.32	4.19	6.02	3.64
BGE-base	Type	776	3.78	12.28	17.45	10.25
BGE-base	Shuffled	776	2.73	8.85	12.84	7.34
E5-base	Depth/lex.	776	N/A	N/A	N/A	N/A
E5-base	Distance	776	0.29	2.82	4.11	2.69
E5-base	Type+dist.	776	0.29	2.64	3.97	2.51
E5-base	Type	776	0.77	7.17	10.86	6.76
E5-base	Shuffled	776	0.55	5.14	7.97	4.84
MPNet	Depth/lex.	776	N/A	N/A	N/A	N/A
MPNet	Distance	776	3.18	6.82	9.24	5.49
MPNet	Type+dist.	776	3.13	6.06	8.66	4.82
MPNet	Type	776	9.80	19.37	25.96	14.34
MPNet	Shuffled	776	7.61	14.72	20.32	10.85
MiniLM	Depth/lex.	776	N/A	N/A	N/A	N/A
MiniLM	Distance	776	2.82	5.86	7.81	4.63
MiniLM	Type+dist.	776	3.32	6.13	8.55	4.87
MiniLM	Type	776	9.97	18.88	25.16	13.63
MiniLM	Shuffled	776	7.67	14.30	19.45	10.21

Table 9: Full ConceptNet false-middle controls. Values are path-level aggregate angular advantages in degrees. ConceptNet is included as a sparse external commonsense-relation check.

Model	Control	n	Δ_C°	Δ_M°	Δ_E°	Euc. Δ_M°
BGE-base	Depth/lex.	1100	N/A	N/A	N/A	N/A
BGE-base	Distance	1100	5.43	12.59	13.36	9.81
BGE-base	Type+dist.	1100	0.02	0.19	0.84	0.19
BGE-base	Type	1100	4.82	11.94	12.83	9.40
BGE-base	Shuffled	1100	4.32	10.81	11.75	8.55
E5-base	Depth/lex.	1100	N/A	N/A	N/A	N/A
E5-base	Distance	1100	1.34	7.89	9.10	7.38
E5-base	Type+dist.	1100	0.04	0.36	0.72	0.35
E5-base	Type	1100	1.03	6.62	7.69	6.26
E5-base	Shuffled	1100	0.89	5.82	6.83	5.52
MPNet	Depth/lex.	1100	N/A	N/A	N/A	N/A
MPNet	Distance	1100	10.01	17.30	17.37	11.55
MPNet	Type+dist.	1100	0.54	1.30	1.90	0.85
MPNet	Type	1100	9.83	18.16	18.27	11.88
MPNet	Shuffled	1100	7.80	14.80	15.24	9.72
MiniLM	Depth/lex.	1100	N/A	N/A	N/A	N/A
MiniLM	Distance	1100	8.81	17.00	16.16	11.81
MiniLM	Type+dist.	1100	0.54	1.28	1.90	0.96
MiniLM	Type	1100	5.68	12.44	12.12	8.99
MiniLM	Shuffled	1100	4.54	10.06	10.21	7.34

Table 10: TREC auxiliary hierarchy-control results. TREC is used only as a hierarchy sanity check, not as evidence for typed lexical relation composition. Values are path-level aggregate angular advantages in degrees.

Score	MiniLM	MPNet	E5-base	BGE-base
<i>Hyp. source</i>				
Cos.	0.50/0.50	0.50/0.50	0.50/0.50	0.50/0.50
R-asym.	0.88/0.88	0.91/0.91	0.92/0.92	0.92/0.92
E-asym.	0.89/0.89	0.91/0.91	0.92/0.92	0.91/0.92
Pot.	0.89/0.89	0.91/0.90	0.91/0.92	0.91/0.90
Struct.	0.82/0.83	0.85/0.86	0.86/0.87	0.85/0.86
<i>Hyp. lexname</i>				
Cos.	0.50/0.50	0.50/0.50	0.50/0.50	0.50/0.50
R-asym.	0.74/0.73	0.85/0.85	0.88/0.88	0.86/0.86
E-asym.	0.81/0.82	0.89/0.89	0.89/0.89	0.87/0.87
Pot.	0.74/0.72	0.78/0.76	0.78/0.77	0.77/0.76
Struct.	0.70/0.74	0.75/0.78	0.75/0.79	0.73/0.76
<i>Mer. source</i>				
Cos.	0.50/0.50	0.50/0.50	0.50/0.50	0.50/0.50
R-asym.	0.79/0.78	0.82/0.81	0.89/0.89	0.86/0.86
E-asym.	0.83/0.83	0.85/0.85	0.89/0.89	0.88/0.88
Pot.	0.85/0.85	0.87/0.87	0.89/0.89	0.86/0.85
Struct.	0.79/0.78	0.83/0.82	0.85/0.84	0.82/0.80
<i>Mer. lexname</i>				
Cos.	0.50/0.50	0.50/0.50	0.50/0.50	0.50/0.50
R-asym.	0.70/0.69	0.74/0.74	0.85/0.85	0.78/0.78
E-asym.	0.78/0.76	0.82/0.81	0.86/0.86	0.85/0.84
Pot.	0.70/0.69	0.66/0.64	0.69/0.68	0.68/0.65
Struct.	0.70/0.69	0.68/0.66	0.70/0.70	0.69/0.67

Table 11: Directionality audit. Each cell reports AUC/AP. Cosine is symmetric and remains at chance, while directional scores recover asymmetric relation structure.

not distinguish the two directions and is expected to remain at chance. Directional scores, including Euclidean and Riemannian asymmetry and scalar potential, can exceed chance if the frozen embedding space preserves asymmetric relation structure. This is shown in Table 11.

A.3 Relation specificity

Directionality alone does not establish that the geometry is relation-specific. A score could distinguish edge direction while still reflecting a generic relatedness or frequency signal. We therefore compare correct relation fields with wrong relation fields. For example, a hypernymy field should score hypernym edges better than a meronymy field does, and vice versa. Table 12 demonstrates this.

A.4 Full HyperLex transfer results

The main text reports only cosine similarity and the best geometric score for each encoder. Table 13 provides the complete method comparison on HyperLex. HyperLex contains graded human lexical-entailment judgements, so it tests whether WordNet-derived relation geometry transfers beyond the diagnostic path task.

Model	n	Correct AUC	Wrong AUC	Gap
<i>Hyp. source</i>				
MiniLM	7825	0.74	0.45	0.28
MPNet	7825	0.76	0.51	0.25
E5-base	7825	0.77	0.63	0.14
BGE-base	7825	0.77	0.57	0.20
<i>Hyp. lexname</i>				
MiniLM	9660	0.62	0.51	0.11
MPNet	9660	0.69	0.53	0.15
E5-base	9660	0.72	0.63	0.09
BGE-base	9660	0.70	0.60	0.11
<i>Mer. source</i>				
MiniLM	5953	0.68	0.49	0.19
MPNet	5953	0.69	0.54	0.15
E5-base	5953	0.74	0.57	0.16
BGE-base	5953	0.73	0.57	0.16
<i>Mer. lexname</i>				
MiniLM	9103	0.63	0.46	0.17
MPNet	9103	0.64	0.50	0.14
E5-base	9103	0.70	0.57	0.13
BGE-base	9103	0.66	0.53	0.13

Table 12: Relation specificity. Correct relation fields outperform wrong relation fields, indicating that the relation geometry is typed rather than merely generic relatedness.

Method	MiniLM	MPNet	E5-base	BGE-base
cosine	0.19/0.19	0.21/0.22	0.19/0.21	0.25/0.27
riem field	0.36/0.35	0.37/0.36	0.38/0.38	0.37/0.35
riem asymmetry	0.34/0.34	0.35/0.35	0.38/0.38	0.37/0.36
euc field	0.33/0.33	0.34/0.33	0.38/0.38	0.37/0.36
euc asymmetry	0.33/0.33	0.34/0.33	0.38/0.38	0.37/0.36
potential	0.32/0.31	0.27/0.26	0.27/0.26	0.36/0.35
lowrank cone	0.38/0.37	0.39/0.38	0.39/0.39	0.39/0.38
structured combined	0.41/0.39	0.38/0.37	0.37/0.36	0.47/0.46

Table 13: Full HyperLex transfer results. Each cell reports Spearman/Pearson correlation with graded lexical-entailment judgements.

Method	MiniLM	MPNet	E5-base	BGE-base
<i>Hyp. source</i>				
cosine	0.73/0.40	0.74/0.40	0.75/0.41	0.74/0.41
riem field	0.74/0.38	0.76/0.41	0.77/0.48	0.77/0.45
riem asymmetry	0.74/0.39	0.76/0.43	0.77/0.48	0.77/0.46
euc field	0.75/0.41	0.76/0.44	0.77/0.48	0.76/0.47
euc asymmetry	0.75/0.41	0.76/0.44	0.77/0.48	0.76/0.47
potential	0.74/0.39	0.75/0.41	0.75/0.42	0.74/0.41
lowrank cone	0.76/0.41	0.78/0.45	0.78/0.50	0.79/0.50
structured combined	0.83/0.63	0.85/0.67	0.85/0.69	0.84/0.67
<i>Hyp. lexname</i>				
cosine	0.71/0.38	0.72/0.38	0.73/0.40	0.73/0.40
riem field	0.62/0.28	0.69/0.32	0.72/0.40	0.70/0.36
riem asymmetry	0.62/0.27	0.69/0.32	0.72/0.40	0.70/0.36
euc field	0.67/0.31	0.72/0.36	0.73/0.41	0.72/0.39
euc asymmetry	0.67/0.31	0.72/0.36	0.73/0.41	0.72/0.39
potential	0.61/0.26	0.63/0.27	0.64/0.28	0.63/0.27
lowrank cone	0.63/0.29	0.70/0.33	0.73/0.41	0.72/0.39
structured combined	0.71/0.48	0.74/0.49	0.74/0.50	0.73/0.48
<i>Mer. source</i>				
cosine	0.69/0.37	0.70/0.37	0.72/0.40	0.71/0.40
riem field	0.68/0.36	0.69/0.36	0.74/0.41	0.73/0.40
riem asymmetry	0.69/0.37	0.70/0.37	0.74/0.41	0.73/0.41
euc field	0.71/0.39	0.71/0.39	0.74/0.42	0.73/0.43
euc asymmetry	0.71/0.39	0.71/0.39	0.74/0.42	0.73/0.43
potential	0.71/0.37	0.72/0.38	0.73/0.38	0.72/0.37
lowrank cone	0.69/0.37	0.70/0.37	0.75/0.44	0.75/0.44
structured combined	0.78/0.53	0.79/0.53	0.81/0.58	0.80/0.54
<i>Mer. lexname</i>				
cosine	0.69/0.37	0.69/0.37	0.73/0.40	0.71/0.39
riem field	0.63/0.29	0.64/0.30	0.70/0.36	0.66/0.31
riem asymmetry	0.63/0.30	0.64/0.31	0.70/0.36	0.66/0.32
euc field	0.67/0.33	0.68/0.34	0.71/0.36	0.70/0.35
euc asymmetry	0.67/0.33	0.68/0.34	0.71/0.36	0.70/0.35
potential	0.62/0.28	0.58/0.26	0.61/0.27	0.60/0.26
lowrank cone	0.65/0.30	0.66/0.31	0.72/0.37	0.69/0.34
structured combined	0.70/0.39	0.67/0.35	0.70/0.38	0.69/0.36

Table 14: Hard-control binary classification. Each cell reports AUC/AP. Structured scores help most in source-disjoint settings, while cosine remains competitive under lexname-held-out domain shift.

A.5 Hard-control binary classification

The path experiments compare real and false middles in two-hop paths. We also evaluate binary relation classification under difficult negative controls. The hard-control setting includes related but incorrect alternatives, such as reversed, sibling or co-part, depth-matched, lexname-matched, and depth-plus-lexname-matched negatives. These controls are designed to reduce the advantage of simple semantic relatedness. Table 14 shows the hard-control binary classification.

A.6 MNLI/HANS sentence-level entailment boundary

The lexical relation results should not be interpreted as evidence of a general sentence-level entailment geometry. To test this boundary, we train sentence-level entailment scores on MNLI and evaluate transfer to HANS. MNLI provides entailment,

Method	MiniLM	MPNet	E5-base	BGE-base
<i>MNLI</i>				
cosine	0.73/0.57	0.81/0.68	0.78/0.64	0.81/0.68
global field	0.50/0.34	0.45/0.31	0.59/0.49	0.57/0.43
global asymmetry	0.51/0.35	0.47/0.32	0.59/0.49	0.57/0.42
<i>HANS</i>				
cosine	0.52/0.49	0.58/0.53	0.57/0.52	0.56/0.51
global field	0.49/0.50	0.50/0.50	0.44/0.47	0.46/0.46
global asymmetry	0.49/0.50	0.51/0.51	0.44/0.47	0.46/0.46

Table 15: Full MNLI/HANS boundary results. Each cell reports AUC/AP. The weak transfer to HANS supports the claim that the observed geometry is lexical and distributional, not a general sentence-level entailment calculus.

Type	Item	Count
edge	AtLocation	1500
edge	CapableOf	1500
edge	Causes	1500
path	AtLocation⇒AtLocation	542
path	AtLocation⇒CapableOf	162
path	Causes⇒CapableOf	49
path	CapableOf⇒CapableOf	7
path	AtLocation⇒Causes	6
path	Causes⇒AtLocation	5
path	Causes⇒Causes	5

Table 16: ConceptNet extraction and path coverage.

neutral, and contradiction labels, whereas HANS is designed to expose failures caused by lexical-overlap, subsequence, and constituent heuristics. Table 15 shows the full MNLI/HANS boundary results.

A.7 ConceptNet coverage and null comparison

ConceptNet is used as an external commonsense-relation resource. Its typed path coverage is uneven after relation filtering, so it is treated as limited external support rather than as a second main benchmark. Table 16 reports the extracted edge and path counts. Table 17 compares semantic embeddings with permuted and random-unit null embeddings under relation-type-distance controls.

A.8 Gloss-mode ablation

The main experiments use lemma-neutral templates to reduce definitional leakage. Gloss-mode embeddings provide a sense-rich upper-bound ablation because they include definitional information. This setting is therefore useful for robustness, but it is not the primary evidence for the paper’s claim. Table 18 shows the gloss-mode WordNet ablation under distance and relation-type-distance controls.

Model	Embedding	n	Δ_C°	Δ_M°	Δ_E°
BGE-base	permuted null	776	0.17	0.27	0.38
BGE-base	random unit null	776	0.04	0.01	0.05
BGE-base	semantic	776	1.32	4.19	6.02
E5-base	permuted null	776	0.04	0.19	0.28
E5-base	random unit null	776	0.04	0.01	0.05
E5-base	semantic	776	0.29	2.64	3.97
MPNet	permuted null	776	0.15	0.04	0.11
MPNet	random unit null	776	0.04	0.01	0.05
MPNet	semantic	776	3.13	6.06	8.66
MiniLM	permuted null	776	0.13	0.04	0.11
MiniLM	random unit null	776	0.07	0.09	0.10
MiniLM	semantic	776	3.32	6.13	8.55

Table 17: ConceptNet semantic versus null embeddings under relation-type-distance controls. Values are path-level aggregate angular advantages in degrees.

Model	Control	n	Δ_C°	Δ_M°	Δ_E°
$H \Rightarrow H$					
MiniLM	Distance	1100	5.07	10.73	15.45
MiniLM	Type+dist.	1100	3.43	8.02	11.62
MPNet	Distance	1100	4.45	10.26	15.00
MPNet	Type+dist.	1100	2.90	7.39	10.88
E5-base	Distance	1100	0.54	4.31	6.97
E5-base	Type+dist.	1100	0.34	3.11	5.10
BGE-base	Distance	1100	2.62	8.37	12.82
BGE-base	Type+dist.	1100	1.69	5.92	9.28
$H \Rightarrow M$					
MiniLM	Distance	1100	4.47	10.20	14.63
MiniLM	Type+dist.	1100	3.65	8.46	12.23
MPNet	Distance	1100	3.89	9.39	13.60
MPNet	Type+dist.	1100	2.89	7.25	10.66
E5-base	Distance	1100	0.47	3.78	6.08
E5-base	Type+dist.	1100	0.38	3.19	5.12
BGE-base	Distance	1100	2.58	7.98	12.16
BGE-base	Type+dist.	1100	1.95	6.20	9.48
$M \Rightarrow H$					
MiniLM	Distance	1100	3.69	8.65	11.51
MiniLM	Type+dist.	1100	3.00	7.32	9.59
MPNet	Distance	1100	3.07	7.66	10.35
MPNet	Type+dist.	1100	2.48	6.47	8.58
E5-base	Distance	1100	0.40	3.53	5.32
E5-base	Type+dist.	1100	0.32	2.96	4.40
BGE-base	Distance	1100	1.94	6.47	9.17
BGE-base	Type+dist.	1100	1.54	5.43	7.57
$M \Rightarrow M$					
MiniLM	Distance	1100	4.15	9.63	13.25
MiniLM	Type+dist.	1100	2.74	7.15	9.60
MPNet	Distance	1100	3.50	8.64	12.18
MPNet	Type+dist.	1100	2.20	6.17	8.43
E5-base	Distance	1100	0.44	3.76	5.87
E5-base	Type+dist.	1100	0.28	2.79	4.29
BGE-base	Distance	1100	2.27	7.27	10.73
BGE-base	Type+dist.	1100	1.38	5.11	7.44

Table 18: Gloss-mode WordNet ablation under distance and relation-type-distance controls. Values are angular advantages in degrees.

Source	ρ_M	ρ_E	Rows
WordNet	0.995	0.998	96
ConceptNet	0.995	0.999	168
TREC	0.989	0.998	24
All	0.995	0.999	288

Table 19: Agreement between Riemannian and Euclidean path metrics, where $\rho_M = \text{Corr}(\Delta_M^R, \Delta_M^E)$ and $\rho_E = \text{Corr}(\Delta_E^R, \Delta_E^E)$. High correlations indicate that the qualitative conclusions are not artefacts of one implementation, while the spherical formulation remains the principled choice for cosine-normalised embeddings.

A.9 Euclidean versus Riemannian geometry

The main formulation uses spherical/Riemannian geometry because the embeddings are normalised and compared using cosine similarity. We nevertheless compare Euclidean and Riemannian variants to test whether the empirical conclusions depend on this implementation choice. The path metrics agreement is shown in Table 19.

A.10 Robustness beyond two-hop paths

The main experiments use two-hop paths because this is the minimal setting in which ordered relation composition can be evaluated while keeping the construction and controls interpretable. As a supplementary robustness analysis, we test whether the true-middle advantage remains positive for three- and four-hop WordNet paths.

For a path of length $L \in \{3, 4\}$, we evaluate 4,400 paths per encoder, covering all 2^L ordered compositions of hypernymy (H) and part-to-whole (M) relations: eight compositions at three hops and sixteen at four hops. The control replaces the complete sequence of internal concepts with a sequence drawn from another path having the same ordered relation sequence. It matches the segment-distance profile and excludes alternatives that form an attested path with the same endpoints and ordered relation types. We use the same seed and bootstrap procedure as in the main two-hop experiment.

For longer paths, closure remains a single endpoint- reconstruction error after composing all edge movements. Mediacy is averaged over the internal concepts, while excess is the total path length minus the direct endpoint distance. Consequently, the three metrics scale differently with path length, and their raw magnitudes should not be compared directly across lengths.

All three advantages are positive for every evaluated encoder-composition group: 32/32 groups

Hops	Paths/ enc.	Comp.	Δ_C°	Δ_M°	Δ_E°
3	4,400	8	2.61	5.11	9.26
4	4,400	16	3.12	4.81	10.23

Table 20: Supplementary longer-path WordNet results under ordered-relation and segment-distance-matched controls. Values are mean angular advantages in degrees over the evaluated encoder-path assignments. ‘‘Comp.’’ denotes the number of ordered relation compositions. Positive values indicate that the resource-defined internal sequence is more coherent than its matched replacement.

Metric pair	True paths	Advantages	Random triples
Mediacy-excess	0.950	0.966	0.610
Closure-mediacy	0.910	0.933	0.613
Closure-excess	0.800	0.860	0.689

Table 21: Pearson correlations among the path metrics for MiniLM. The semantic-path correlations exceed those arising from random spherical geometry, but the metrics remain substantially dependent.

at three hops and 64/64 groups at four hops. This result indicates that the true-path advantage is not unique to the minimal two-hop construction. It should not, however, be interpreted as showing that coherence improves or deteriorates monotonically with path length, because the metrics and control problem scale differently as paths become longer.

A.11 Dependence among path metrics

Closure, mediacy, and excess are not designed to provide independent statistical tests. All three are functions of the same points on the unit sphere, and some correlation is therefore expected even for random spherical triangles. Table 21 reports Pearson correlations for MiniLM under the relation-type-and-distance-matched control, together with a baseline of 100,000 random triples on a 384-dimensional unit sphere.

These results support interpreting the metrics as complementary views of a shared path-coherence structure, rather than as three independent confirmations. Closure nevertheless has a distinct operational interpretation because it evaluates endpoint reconstruction after parallel transport and composition.

A.12 Artefact licences and intended use

We use Princeton WordNet 3.0 under the WordNet License and ConceptNet 5 under CC BY-SA 4.0.

The Brown Corpus distributed through NLTK is used under its non-commercial-use terms. HANS is released under the MIT License, while MultiNLI is used under the source-specific licensing and permission terms documented with its release.

The all-MiniLM-L6-v2 and all-mpnet-base-v2 model releases use the Apache License 2.0; e5-base-v2 and bge-base-en-v1.5 use the MIT License. HyperLex and the Li-Roth question-classification data are used for non-commercial research evaluation and cited as requested by their official distributions, which do not state a separate standard dataset licence.

All artefacts are used consistently with their intended research and evaluation purposes: pre-trained encoders for representation analysis, lexical resources for constructing relation paths, benchmark datasets for evaluation, and software packages for implementation and statistical analysis. We do not redistribute raw datasets or pretrained model weights.