When Structure Matters: Cross-Lingual Hyperbolic Embeddings for Chinese and English Wordnets

Mao-Chang Ku, Da-Chen Lian, Pin-Er Chen, Po-Ya Angela Wang, Wei-Ling Chen, Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University, Taiwan
d08142002@ntu.edu.tw, d08944019@ntu.edu.tw, cckk2913@gmail.com,
differe94nt@gmail.com, d10142007@ntu.edu.tw, shukaihsieh@ntu.edu.tw

Abstract

Hyperbolic embeddings such as the Poincaré model effectively represent lexical hierarchies with low distortion, yet their cross-lingual generalizability remains largely unexplored. This study investigates cross-lingual transfer by training 20-dimensional Poincaré embeddings exclusively on Open English WordNet (OEWN) hypernymy relations and evaluating on aligned Chinese Wordnet (CWN) synsets under a vocabulary-constrained transfer setting, where CWN-relevant synsets appear in OEWN training data but no Chinese-language supervision is used. We report robust statistical evidence based on the final 10 training checkpoints: Poincaré embeddings achieve 2.57× higher Mean Reciprocal Rank (MRR) than Euclidean embeddings on CWN (0.030 $\pm$ 0.001 vs 0.012 $\pm$ 0.000, $p < 0.001$, Cohen's $d = 34.48$) and 5.61× higher on OEWN (0.016 $\pm$ 0.000 vs 0.003 $\pm$ 0.000, $p < 0.001$, $d = 42.48$). Furthermore, hierarchical filtering leveraging the radial dimension of hyperbolic space provides substantial additional gains: +74.6% MRR improvement on CWN and +25.8% on OEWN (both $p < 0.001$). The model achieves higher absolute performance on the zero-shot CWN test set (MRR = 0.052 $\pm$ 0.002) than on the in-domain OEWN test set (MRR = 0.020 $\pm$ 0.001). We attribute this to structural alignment: CWN's broader branching factor (4.32 vs 1.10) and moderate depth naturally suit hyperbolic geometry's capacity to compactly represent hierarchies. Our findings demonstrate that geometric properties learned from English hypernymy transfer robustly across languages when semantic structures align. We release the aligned CWN–OEWN hypernymy evaluation dataset and complete evaluation framework to facilitate future research on geometry-based cross-lingual semantic modeling.

Keywords: Hyperbolic embeddings, Chinese Wordnet, hypernymy relations

1. Introduction

Modeling hierarchical semantic relations in lexical networks is a key challenge in computational semantics. Euclidean embeddings capture similarity but struggle with representing hierarchical depth and branching (De Sa et al., 2018; Nickel and Kiela, 2017). Hyperbolic geometry, with its exponential volume growth, enables compact and low-distortion representations of trees and taxonomies. The Poincaré embedding framework effectively modeled WordNet hypernymy by mapping semantic generality to proximity to the origin (Nickel and Kiela, 2017). Later studies expanded hyperbolic representation learning to various manifolds and asymmetric relations, consistently outperforming Euclidean spaces in preserving taxonomic order and hierarchy depth (Nickel and Kiela, 2018; Ganea et al., 2018; Le et al., 2019). Although recent work explored its potential for cross-lingual representation (Saxena et al., 2022), research remains largely confined to English WordNet. Unlike English WordNet, which exhibits deep taxonomic chains reflecting Western categorical traditions, Chinese Wordnet reflects a conceptual organization that prioritizes functional generality and broader semantic groupings over fine-grained categorical depth. For instance, the concept *piano* is organized under a deep chain in English WordNet (*musical instrument* > *device* > *artifact* > *entity*), whereas in CWN it connects directly to broad categories such as "musical instrument" and "equipment"—a pattern typical of CWN's broader, shallower taxonomy. This structural difference provides an ideal testbed for examining whether geometric properties of hierarchical embeddings transfer across typologically distinct lexical organizations. This structural difference provides an ideal testbed for examining whether geometric properties of hierarchical embeddings transfer across typologically distinct lexical organizations.

This study aims to extend hyperbolic embeddings to cross-lingual lexical networks by training on hypernymy relations from Princeton WordNet (PWN) (Miller, 1995) and projecting corresponding Chinese Wordnet (hereafter, CWN) nodes (Huang et al., 2010) into a shared space. We evaluate how well this space preserves CWN's hierarchical relations, assessing the cross-lingual generalizability of hyperbolic modeling and providing new insights into the structural alignment of multilingual lexical resources.

This study makes several contributions to cross-lingual hierarchical representation learning. We demonstrate robust cross-lingual transfer of hyperbolic embeddings from English to Chinese without language-specific training, achieving higher per-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 12054–12071
11–16 May 2026.

formance on zero-shot Chinese evaluation than on in-domain English tests. We provide both theoretical and empirical evidence linking hierarchical structure—specifically branching factor and depth—to embedding geometry effectiveness, explaining why structural alignment matters more than language-specific patterns. To support reproducibility and future research, we release the CWN–OEWN aligned hypernymy dataset and complete evaluation framework for cross-lingual hierarchy modeling.

The rest of the paper is organized as follows. We briefly reviews related work, describes the embedding framework and dataset construction (Section 3), illustrate the experimental results (Section 4), and finally concludes with limitations and future directions.

2. Related Work

2.1. Hyperbolic Embeddings

Hyperbolic embeddings provide an efficient framework for modeling hierarchical relations in lexical semantics. Nickel and Kiela (2017) introduced the Poincaré embeddings, mapping WordNet synsets into a hyperbolic space that preserves hierarchical depth with substantially lower distortion than Euclidean embeddings. De Sa et al. (2018) analyzed the theoretical trade-offs of such models, showing why hyperbolic spaces encode hierarchies compactly in low dimensions. The Lorentz model (Nickel and Kiela, 2018) refined the formulation by improving numerical stability, optimization efficiency, and scalability for large datasets. However, Bansal and Benton (2021) found that high-dimensional Euclidean embeddings can reach comparable accuracy, suggesting that hyperbolic geometry's strength lies mainly in low-dimensional efficiency.

Building on these foundations, later studies extended hyperbolic representation learning to model complex semantic and relational structures. Ganea et al. (2018) modeled asymmetric hierarchical relations through entailment cones in hyperbolic space, where inclusion between cones encodes hypernymy and hyponymy. Balažević et al. (2019) proposed the Multi-Relational Poincaré (MuRP) model to represent overlapping hierarchies in knowledge graphs, while Le et al. (2019) demonstrated that hierarchical relations can be automatically induced from text using Hearst patterns under hyperbolic constraints. Finally, Saxena et al. (2022) aligned cross-lingual hierarchies within shared hyperbolic spaces, suggesting their potential for representing lexical networks in other languages.

2.2. Chinese Wordnet

Chinese Wordnet (CWN) extends the WordNet framework to Mandarin Chinese, representing lexical semantics via synsets and explicit semantic relations. Adopting PWN's ontology but adapted for Mandarin Chinese, CWN accounts for language-specific lexicalization, part-of-speech distinctions, and sense granularity across nouns, verbs, adjectives, and adverbs. The latest release[1] of CWN includes 29,321 lexical entries, 12,620 synsets, and nearly 60,000 semantic relations, preserving WordNet's architecture while reflecting Mandarin typology, such as morphological productivity and polysemy. It provides a linguistically grounded resource for examining how Mandarin Chinese semantic hierarchies can be embedded in hyperbolic space.

3. Methodology

3.1. Poincaré Embeddings

Poincaré embeddings (Nickel and Kiela, 2017) model hierarchical relations by mapping entities into an n-dimensional Poincaré ball, a Riemannian manifold with constant negative curvature. The exponential growth of distances from the origin allows compact representation of tree-like structures.

Formally, each node x_i is represented as a point in the open unit ball:

$$\mathbb{B}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}. \tag{1}$$

The distance between two points u and v in this ball is defined by the Poincaré metric:

$$d_{\mathbb{B}}(u,v) = \operatorname{arcosh}\left(1 + 2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|v\|^2)}\right). \tag{2}$$

To learn embeddings, the model minimizes a ranking-based loss to preserve observed hierarchical relations while separating unrelated pairs:

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d_{\mathbb{B}}(u,v)}}{\sum_{v'\in N(u)} e^{-d_{\mathbb{B}}(u,v')}}, \tag{3}$$

where $\mathcal{D}$ denotes true hypernym–hyponym pairs and $N(u)$ represents negative samples. Optimization is performed via Riemannian stochastic gradient descent to ensure all updates remain within the unit ball. This implementation follows Nickel and Kiela (2017), which trained on WordNet hypernymy relations to obtain low-distortion embeddings of lexical hierarchies.

We use the aforementioned implementation with modifications to support a two-stage learning rate

[1] https://lopentu.github.io/CwnWeb/

schedule. We train 20-dimensional embeddings for 3000 epochs (see the Appendix for complete training details).

To validate the effectiveness of hyperbolic geometry for hierarchical representation, we train Euclidean embeddings as a baseline for comparison, using the same model architecture with appropriately scaled learning rates (0.5 and 0.25 for the two training stages). This comparison follows established practice in prior work (Nickel and Kiela, 2017; Sala et al., 2018), where Poincaré embeddings have consistently demonstrated superior performance over Euclidean baselines for hierarchical data, as shown in Section 4.2.

Our implementation extends the Nickel and Kiela (2017) and Bansal and Benton (2021) codebases to support two-stage learning rate scheduling and checkpoint saving. To ensure full reproducibility, we will make all code, preprocessing scripts, and trained models publicly available upon acceptance.

3.2. Data

We use the `wn` Python package[2] (Goodman and Bond, 2021) to extract the noun hypernymy relations from PWN. Similarily, we use the `CwnGraph 0.3.0`[3] to extract noun hypernymy relations from CWN.

Since our hyperbolic embedding space is trained on English data, testing whether the hierarchical relations in CWN can be captured within this space requires extracting the CWN synsets that can be mapped to corresponding synsets in the PWN.[4] We thus align the CWN synsets with PWN synsets by instructing `gpt-5-mini`. The instruction prompt is shown in Appendix A.

3.3. Dataset Split Strategy

We term our evaluation setting *vocabulary-constrained transfer*: no Chinese-language training is used, but the protected-node strategy (described below) ensures all CWN test synsets have corresponding representations in the OEWN-trained embedding space, preventing out-of-vocabulary errors. While this departs from a strictly zero-shot protocol, it isolates the effect of geometric cross-lingual transfer from vocabulary coverage issues. We discuss this limitation further in Section 5.

Creating a valid train-test split for hierarchical relation prediction presents a critical challenge: ensuring vocabulary integrity across splits. Unlike traditional classification tasks, where samples can be randomly partitioned, our graph-structured data requires that all nodes (synsets) appearing in test edges must also appear in the training set to prevent out-of-vocabulary (OOV) errors during evaluation.

We employed a vocabulary-preserving split strategy on the OEWN noun hypernymy graph, which contains 88,200 hyponym-hypernym pairs. To ensure comparability with the CWN test set (1,182 edges), we targeted a test set size of approximately 2% of the OEWN data (1,764 edges).

Our splitting procedure follows these steps:

1. **Identify protected edges**: We first identified 12,348 edges containing at least one synset that appears in our cross-lingual CWN evaluation set. These edges are guaranteed to remain in the training set to ensure the model learns representations for all CWN-relevant synsets.

2. **Sample candidate test edges**: From the remaining 75,852 edges connecting synsets not directly involved in CWN evaluation, we randomly sampled candidate test edges.

3. **Validate vocabulary coverage**: Each candidate test edge was validated to ensure both its source (hyponym) and target (hypernym) nodes appear in at least one training edge. Edges failing this criterion were rejected to prevent OOV errors.

4. **Construct final splits**: After validation, 499 edges met the vocabulary coverage requirement, yielding a final training set of 87,701 edges. All test edges are guaranteed to contain only synsets present in the training vocabulary.

This conservative approach prioritizes evaluation validity over test set size.

3.4. Test Sets

We employ two test set formats to enable comprehensive evaluation across different aspects of hierarchical modeling.

Pairwise test sets (1,182 CWN pairs, 499 OEWN pairs) contain hyponym–hypernym pairs for evaluating single-answer link prediction. For each pair (u, v), we rank the true hypernym u against all other synsets in the vocabulary. We

[2]We use the 0.13.0 version of `wn` package (i.e., oewn:2024), the Open English Wordnet derived from PWN.

[3]https://github.com/lopentu/CwnGraph/tree/develop

[4]For example, if CWN synset α corresponds to PWN synset A, and CWN synset β corresponds to PWN synset B, and α is a hypernym of β, then we use the A–B pair as a test case. If the model can correctly capture the relation between A and B, this indicates its ability to recognize the corresponding hypernymy relation between α and β.

compute both *hierarchical ranking* (incorporating norm-based constraints where valid hypernyms must have smaller norms than their hyponyms) and *distance-only ranking* (using Poincaré distance alone).

Hierarchical test sets (997 CWN synsets, 498 OEWN synsets) contain complete ancestor structures organized by graph distance level (level 1 = immediate parents, level 2 = grandparents, etc.). Each entry includes all ancestors at each hierarchical level, enabling multi-answer ranking evaluation and geometric fidelity assessment.

3.5. Evaluation Metrics

We evaluate model performance using complementary metrics that assess different aspects of hierarchical representation learning. All metrics are computed for both OEWN and CWN test sets.

Pairwise Ranking Metrics. Using the pairwise test sets, we evaluate the model's ability to identify the correct hypernym by ranking it against all vocabulary synsets:

- **Mean Reciprocal Rank (MRR):** Measures the average inverse rank of the true hypernym. A higher score indicates the model consistently places correct hypernyms at the top of the ranking. Values range from 0 to 1.

Multi-Answer Ranking Metrics. Using the hierarchical test sets, we evaluate the model's ability to rank multiple valid ancestors for each query synset:

- **Mean Average Precision (MAP):** Evaluates the overall quality of the ranking by computing precision at each position where a true ancestor appears, then averaging across all ground-truth ancestors. Values range from 0 to 1, where higher scores indicate true hypernyms are consistently ranked higher.
- **Hits@K:** Measures the percentage of test cases where at least one true ancestor is found within the top-K predictions. We report Hits@1, Hits@3, and Hits@10 to assess performance at different retrieval depths.
- **Normalized Discounted Cumulative Gain (nDCG@10):** Assesses ranking quality with position-based discounting and level-aware relevance scores. We assign relevance as relevance $= \max(3-(\ell-1), 1)$, prioritizing immediate ancestors (level 1: relevance=3) over distant ones (level ≥ 3: relevance=1), then apply logarithmic position-based discounting. Values range from 0 to 1.

Geometric Fidelity Metrics. Beyond topological accuracy, we assess whether embedding distances accurately reflect hierarchical structure. Using the hierarchical test sets, we compare the **ground-truth hierarchical distance** $d_{\text{true}}(u, v)$ (shortest path length in the Wordnet graph) against the **predicted distance** $d_{\text{pred}}(u, v)$ (Poincaré distance between embeddings, Equation 2):

- **Spearman's Rank Correlation (ρ):** Assesses whether the model preserves the monotonic relationship between graph distance and embedding distance. Values close to 1 indicate that more distant ancestors consistently have larger embedding distances.
- **Mean Absolute Error (MAE):** Measures the average absolute deviation between predicted and true hierarchical levels:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |d_{\text{pred}}(u_i, v_i) - d_{\text{true}}(u_i, v_i)| \quad (4)$$

  Lower MAE indicates more accurate distance preservation.

3.6. Statistical Analysis

To ensure robust and generalizable conclusions, we adopt a rigorous statistical methodology. Rather than reporting single checkpoint results, we analyze the final 10 training checkpoints (epochs 2775–3000; each checkpoint is 25 epochs apart), representing the converged state of training. For each metric, we compute mean and sample standard deviation across these checkpoints.

Statistical significance is assessed using paired t-tests comparing the same checkpoints under different conditions (e.g., Poincaré vs Euclidean). We selected paired t-tests specifically to account for correlated checkpoint measurements and avoid inflated significance that could arise from treating checkpoints as independent samples. This conservative approach ensures our statistical claims remain valid despite multiple measurements from a single training run. We report Cohen's d effect sizes to quantify practical significance.

4. Results and Discussion

4.1. Learned Hierarchical Structure

To understand how Poincaré embeddings organize semantic hierarchies, we first visualize the learned representations. Figures 1 and 2 show the biological taxonomy from the OEWN training data, projected from 20 dimensions to 2D and 3D views, respectively, for visualization.

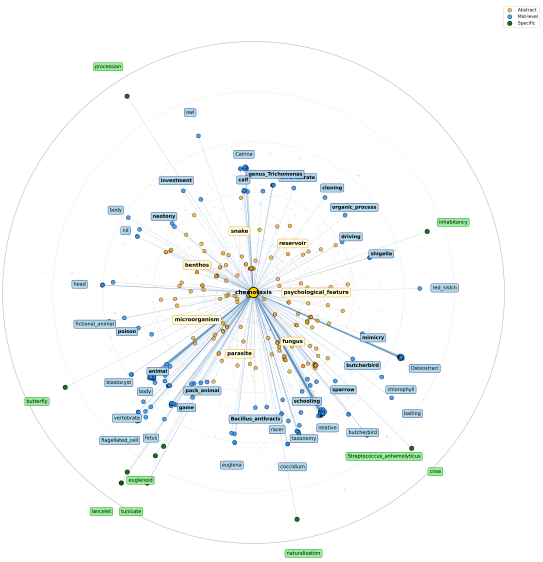


Figure 1: 2D visualization of biological taxonomy in Poincaré space demonstrating cross-lingual structural alignment. The radial gradient reflects learned semantic generality without explicit supervision, with abstract concepts near the center and specific instances toward the periphery.

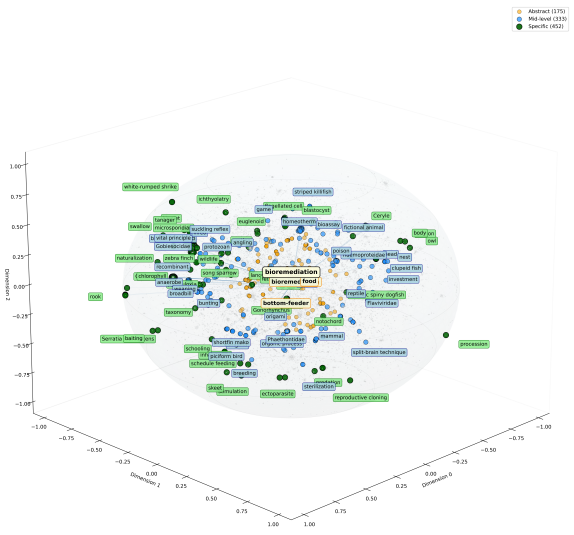


Figure 2: 3D view of biological taxonomy in the Poincaré ball. The radial organization demonstrates how hyperbolic space naturally encodes semantic generality, with abstract concepts near the center and specific instances toward the boundary.

The model exhibits the key theoretical property of hyperbolic embeddings: a clear radial gradient from abstract to specific concepts. Abstract biological categories cluster near the center, mid-level terms occupy intermediate positions, and specific instances position toward the periphery. This three-tier organization spanning 960 biological terms emerges purely from learning hypernymy relations.

4.2. Performance Evaluation

We now evaluate the model's performance using Mean Reciprocal Rank (MRR) and Hits@K for link prediction. Additional ranking metrics (e.g., MAP and nDCG) and hierarchical distance analyses (Spearman's ρ and MAE) are reported in Appendix C for completeness.

We report results based on the final 10 training checkpoints (epochs 2775–3000) to account for training variance, following the statistical methodology described in Section 3.6.

Geometry Comparison. We first compare Poincaré and Euclidean embeddings using distance-only ranking to ensure fair evaluation. Poincaré embeddings significantly outperform Euclidean across all metrics (Table 1, Figures 3 and 4). On the cross-lingual CWN test set, Poincaré achieves 2.57× higher MRR (0.030 $\pm$ 0.001 vs 0.012 $\pm$ 0.000, $t(9) = 95.81$, $p < 0.001$, Cohen's $d = 34.48$), with Poincaré outperforming Euclidean in all 10 checkpoints tested. On the in-domain OEWN test set, the advantage is even larger: 5.61× higher MRR (0.016 $\pm$ 0.000 vs 0.003 $\pm$ 0.000, $p < 0.001$, $d = 42.48$). Similar patterns emerge for Hits@K metrics, with Poincaré achieving 5.8× better Hits@1 performance on both test sets. The extremely large effect sizes (Cohen's $d > 30$) reflect the fundamental geometric difference between hyperbolic and Euclidean spaces for hierarchical data: as demonstrated theoretically by Sarkar (2011) and Sala et al. (2018), hyperbolic embeddings achieve qualitatively lower distortion for tree-like structures regardless of task setting, making large effect sizes expected rather than incidental. We acknowledge, however, that the structural imbalance between our CWN and OEWN test sets (Section 4.3) may amplify the apparent magnitude of these effects; future work with structurally matched test sets would better isolate geometry from structural bias.

Structural Alignment Effects. When using hierarchical filtering (which leverages the radial dimension of the Poincaré ball to filter candidates by semantic generality; in other words, we select synsets with norms smaller than the target synset as potential hypernyms), the model achieves higher absolute performance on the zero-shot CWN test set (MRR = 0.052 $\pm$ 0.002) than on the in-domain OEWN test set (MRR = 0.020 $\pm$ 0.001, $t(9) = 26.70$, $p < 0.001$, $d = 6.78$). This result reflects structural alignment between CWN's topology and hyperbolic geometry's strengths. As

Table 1: Performance comparison between Poincaré and Euclidean embeddings on distance-only ranking metrics. Values are mean ± SD across the final 10 checkpoints. ***$p < 0.001$ (paired t-test, $n = 10$). Bold indicates best performance.

Metric	CWN (Cross-lingual)		OEWN (In-domain)	
	Poincaré	Euclidean	Poincaré	Euclidean
MRR	**0.030 ± 0.001*****	0.012 ± 0.000	**0.016 ± 0.000*****	0.003 ± 0.000
MAP	**0.063 ± 0.001*****	0.020 ± 0.000	**0.027 ± 0.001*****	0.005 ± 0.000
Hits@1 (%)	**15.8 ± 0.4*****	2.7 ± 0.1	**10.4 ± 0.2*****	1.8 ± 0.0
Hits@3 (%)	**35.7 ± 0.3*****	10.5 ± 0.1	**20.2 ± 0.5*****	5.4 ± 0.1
Hits@10 (%)	**72.0 ± 0.5*****	51.1 ± 0.2	**38.0 ± 0.8*****	13.3 ± 0.2
Mean Rank	**9,337 ± 23*****	20,351 ± 28	**9,473 ± 45*****	29,559 ± 32

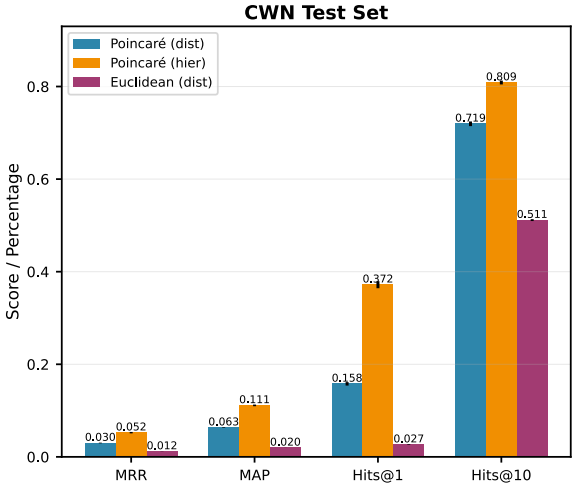


Figure 3: Performance comparison on CWN test set (cross-lingual transfer). Poincaré embeddings (distance-only: blue, hierarchical: orange) outperform Euclidean embeddings (magenta) across all metrics. Error bars represent ±1 SD across final 10 training epochs (2775–3000).

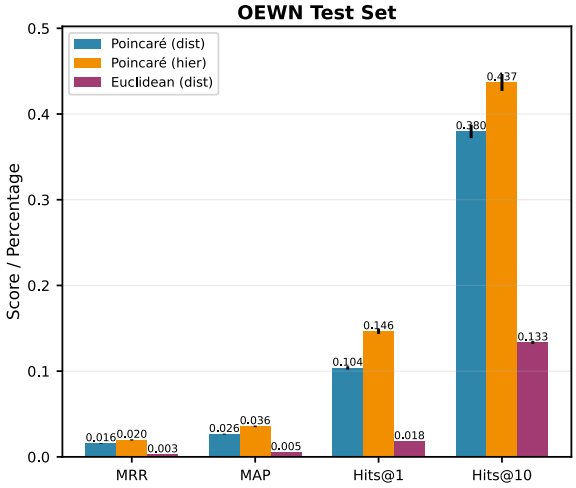


Figure 4: Performance comparison on OEWN test set (in-domain evaluation). Poincaré embeddings show substantial advantages over Euclidean, with hierarchical filtering providing additional improvements. Error bars represent ±1 SD.

detailed in Section 4.3, CWN exhibits substantially broader branching (mean 4.32 vs 1.10) and moderate depth (mean 2.51 vs 1.02) compared to our OEWN test set. Theoretical work demonstrates that hyperbolic embeddings achieve optimal distortion for precisely such structures—broad, shallow hierarchies where many children connect to common parents at similar radial depths (Sarkar, 2011; Sala et al., 2018). In contrast, the OEWN test set's narrow structure (max branching 5, largely single parent-child relations) provides less opportunity for hyperbolic geometry to exploit its exponential capacity.

Hierarchical Filtering Advantage. Beyond geometry comparison, hierarchical filtering provides substantial additional improvements over distance-only ranking for Poincaré embeddings (Table 2). On the CWN test set, hierarchical filtering increases MRR by 74.6% (0.030 → 0.052, $t(9) = 77.14$, $p < 0.001$, $d = 18.81$), with particularly dramatic improvements for Hits@1 (+135.8%: 15.8% → 37.2%). At Hits@10, accuracy reaches 80.9% ± 0.4%, indicating that in over four-fifths of cases, a correct ancestor appears within the top 10 candidates. This suggests that the model effectively locates the correct semantic region of the hierarchy even when it does not identify the exact immediate parent. On the OEWN test set, gains are more modest but still significant: +25.8% for MRR and +41.0% for Hits@1 (both $p < 0.001$). The larger gains on CWN directly reflect its broader branching structure, which provides more opportunities for norm-based filtering to disambiguate among candidates by leveraging the radial coordinate (distance from origin) that encodes semantic generality in hyperbolic space.

4.3. Wordnet Comparison

We now analyze the structural properties that explain the performance differences observed in Sec-

Table 2: Hierarchical filtering advantage for Poincaré embeddings. Values are mean ± SD across final 10 training epochs. ***$p < 0.001$ (paired t-test, $n = 10$). Improvements show percentage gains for MRR/MAP/Hits@K and absolute reduction for Mean Rank.

	CWN (Cross-lingual)			OEWN (In-domain)		
Metric	Distance-Only	Hierarchical	Improv.	Distance-Only	Hierarchical	Improv.
MRR	0.030±0.001	**0.052±0.002*****	+74.6%	0.016±0.000	**0.020±0.001*****	+25.8%
MAP	0.063±0.001	**0.111±0.002*****	+75.9%	0.027±0.001	**0.036±0.001*****	+34.4%
Hits@1 (%)	15.8±0.4	**37.2±0.8*****	+135.8%	10.4±0.2	**14.6±0.3*****	+41.0%
Hits@3 (%)	35.7±0.3	**64.4±0.5*****	+80.4%	20.2±0.5	**26.4±0.5*****	+30.6%
Hits@10 (%)	72.0±0.5	**80.9±0.4*****	+12.4%	38.0±0.8	**43.7±1.0*****	+15.0%
Mean Rank	9,337±23	**2,769±12*****	$-6,568$	9,473±45	**5,573±29*****	$-3,900$

tion 4.2. Table 3 presents a side-by-side comparison of branching factors and hierarchical depth for both test sets.

Within our test sets, CWN displays substantially broader branching (mean 4.32, max 202) compared to OEWN (mean 1.10, max 5). In the CWN test data, a majority of hyponyms reach a root category within 2–3 steps, and none required more than 7 steps. This indicates that Chinese taxonomic chains are generally shorter (shallower) than might be expected in English WordNet when fully considered. For instance, English WordNet often has longer chains where a specific species might go through multiple intermediate classes (genus, family, order, etc.) before reaching a top concept like "organism." Such vertical stratification is largely absent in the CWN test—many Chinese concepts jump directly to a broad umbrella category. For instance, in CWN, "piano" directly connects to broad categories like "musical instrument" and "equipment," whereas English WordNet places it under a deeper chain: "musical instrument > device > artifact > entity."

CWN hypernyms typically exhibit extensive "horizontal" branching—each broad category encompassing many hyponyms. For example, in the CWN test set, the hypernym "unit (of measurement)" covers 202 distinct hyponyms, and "human relationship" includes over 50 hyponym entries (e.g., various kinship terms). By contrast, no English hypernym in the OEWN test set governs more than five hyponyms. This fundamental difference—CWN test data feature broad and yet moderately deep hierarchies (mean branching 4.32; max 202; depth ≈2–3), while OEWN test is narrow and very shallow (branching mean 1.10; max 5; depth ≈1)—has important theoretical implications. The top 15 synsets by number of hyponyms for CWN and OEWN test data are shown in Appendix D.

Theoretical Justification. Hyperbolic embeddings are provably optimal for precisely the structural properties exhibited by the CWN test set. First, Sarkar (2011) shows any finite tree can be embedded into the hyperbolic plane as a Delaunay embedding, preserving parent-child distances exactly up to global scaling, and achieving distance distortion $\leq 1 + \epsilon$ between non-adjacent nodes for arbitrarily small ϵ. Second, Sala et al. (2018) derive precision-dimensionality tradeoffs, proving that hyperbolic embeddings achieve low distortion for hierarchies when branching is high but depth is moderate. They demonstrate that hyperbolic spaces outperform Euclidean embeddings on Wordnet-like structures, especially in low dimensions, because they can represent many children per parent at similar radial depth with manageable distortion.

These theoretical results explain our empirical findings: the CWN test set's broad, shallow structure aligns perfectly with hyperbolic geometry's strengths, enabling the model to achieve higher absolute performance despite zero-shot, cross-lingual transfer. Furthermore, hierarchical filtering exploits this alignment by using the radial coordinate to filter candidates, explaining why the +74.6% MRR improvement on CWN substantially exceeds the +25.8% improvement on OEWN. The broader branching in CWN provides more candidates at each hierarchical level, giving norm-based filtering more opportunities to eliminate semantically distant options.

Linguistic Interpretation. This structural asymmetry between CWN and OEWN reflects deeper typological differences: Mandarin's productive compounding and preference for category-general nouns yield broader but shallower hierarchies, where many specific concepts connect directly to general umbrella terms. English WordNet's narrower, deeper structure reflects a Linnaean taxonomic tradition emphasizing graduated categorical refinement. The superior performance on CWN demonstrates that hyperbolic geometry naturally aligns with Chinese lexical organization, suggesting geometry-structure fit matters more than language-specific surface patterns.

Table 3: Side-by-side comparison of CWN and OEWN test-set branching statistics. *Stats computed on induced test-set graphs.*

Metric	CWN (Native)	CWN ($\rightarrow$OEWN)	OEWN Test
Hypernym–hyponym pairs	1,473	1,473	499
Unique synsets (nodes)	1,481	1,268	942
Branching Mean	4.32	4.06	1.10
Branching Max	202	140	5
Depth Mean	2.51	2.40	1.02
Depth Max	7	7	2

5. Limitations

While our findings demonstrate the potential of hyperbolic embeddings for cross-lingual knowledge transfer, several limitations warrant consideration.

First, the performance difference between OEWN and CWN test sets may be partially attributable to our sampling strategy. The OEWN test set (499 edges) was intentionally constructed to be substantially smaller and structurally simpler than the CWN test set (1,473 edges), with markedly lower branching factors (1.10 vs. 4.32) and depth (1.02 vs. 2.51). This design choice, while ensuring vocabulary integrity, creates an unbalanced comparison that directly confounds cross-lingual transfer effects with structural effects: the performance advantage on CWN may partly reflect its more favorable branching topology rather than successful language transfer per se. Readers should therefore interpret the CWN–OEWN performance gap as an upper-bound estimate of cross-lingual transfer benefit. Future work should employ stratified sampling to construct tructurally matched test sets—equating branching factor and depth distributions across languages—to disentangle these two sources of performance difference.

Second, our cross-lingual evaluation relies on automatic synset alignment using `gpt-5-mini` without manual validation or inter-annotator agreement metrics (e.g., Cohen's κ). This is a substantive limitation: since our core cross-lingual transfer claims depend on this alignment, mapping errors could directly affect reported performance. Culture-specific concepts lacking English equivalents may also be systematically excluded, biasing evaluation toward more universal semantic categories and potentially inflating transfer performance. Future work should include human spot-checking of a stratified alignment sample and report agreement scores to bound the uncertainty introduced by automatic alignment.

Third, our evaluation scope is limited to noun hypernymy relations in a single language pair (English-Chinese) using only the Poincaré model in 20 dimensions. The generalizability to other parts of speech, semantic relations (e.g., meronymy, antonymy), language pairs with greater typological distance, or alternative hyperbolic formulations (e.g., Lorentz model) remains unexplored.

Finally, our evaluation design is not purely zero-shot. The "protected edges" strategy ensures that all synsets in the CWN test set appear in the OEWN training set, meaning the model has been explicitly trained on the English-space representations of test nodes. While this prevents out-of-vocabulary errors, a strictly zero-shot evaluation would require testing on entirely unseen synsets. Future work should also include confidence intervals, statistical significance tests, and fine-grained analysis stratified by branching factor, depth, and semantic domain to provide more robust evidence for the cross-lingual transferability of hyperbolic embeddings.

6. Conclusion

This study examines the cross-lingual transferability of Poincaré embeddings by training a 20-dimensional model on OEWN hypernymy relations and evaluating it zero-shot on aligned CWN synsets. Our results demonstrate both the superiority of hyperbolic over Euclidean geometry and the importance of structural alignment for cross-lingual transfer.

Poincaré embeddings achieve massive advantages over Euclidean baselines, with 2.57$\times$ higher MRR on CWN ($p < 0.001$, Cohen's $d = 34.48$) and 5.61$\times$ higher on OEWN ($p < 0.001$, $d = 42.48$), with these effects consistent across all 10 final training checkpoints analyzed. Hierarchical filtering, which leverages the radial dimension of hyperbolic space to filter by semantic generality, provides substantial additional gains: +74.6% MRR improvement on CWN and +25.8% on OEWN (both $p < 0.001$).

The model achieves higher absolute performance on the zero-shot CWN test set (MRR = 0.052 $\pm$ 0.002) than on the in-domain OEWN test set (MRR = 0.020 $\pm$ 0.001), reflecting structural alignment rather than language-specific ad-

vantages. CWN's topology—characterized by broader branching (mean 4.32 vs 1.10) and moderate depth (mean 2.51 vs 1.02)—matches the structural properties for which hyperbolic embeddings are theoretically optimal (Sarkar, 2011; Sala et al., 2018). This structural alignment enables the model to exploit the exponential capacity of hyperbolic space more effectively on CWN than on OEWN's narrower test structure.

To support reproducibility and facilitate future research, we make available the aligned CWN–OEWN test sets with complete ancestor structures, evaluation scripts for all metrics reported, and preprocessing pipelines for vocabulary-preserving splits. These resources enable benchmarking of alternative embedding methods and extension to other language pairs. Our findings provide empirical evidence for geometry-based multilingual resource integration and pave the way for structure-aware lexical alignment across typologically diverse languages.

Overall, our findings highlight that cross-lingual transfer of hyperbolic embeddings depends critically on structural alignment between source and target hierarchies. When semantic structures align—specifically, when target hierarchies exhibit the broad, shallow branching that hyperbolic geometry efficiently captures—geometric properties learned from one language transfer robustly to another. This suggests that hyperbolic embeddings capture universal principles of hierarchical organization rather than language-specific surface patterns, providing a foundation for multilingual semantic representation.

This work opens several avenues for future research. Extending our approach to typologically distant languages would test whether geometry-structure alignment holds across diverse lexical organizations beyond the English-Chinese pair. Additionally, investigating whether Chinese-origin language models such as Qwen (Bai et al., 2023), which may generate typologically different semantic graphs for Chinese, alter cross-lingual transfer dynamics would test whether geometry-structure fit is robust to the choice of underlying language resource. Incorporating relation types beyond hypernym—including meronymy, antonymy, and causation—could enable development of multi-relational hyperbolic knowledge graphs that capture richer semantic structures. Exploring dynamic embeddings that model diachronic semantic change in multilingual hierarchies would reveal whether geometric properties remain stable across historical language evolution. Finally, developing a standardized multilingual hyperbolic lexical graph benchmark with unified evaluation protocols would facilitate systematic comparison of embedding methods across languages. We envision these extensions contributing to a comprehensive framework for geometry-aware cross-lingual semantic modeling.

7. Bibliographical References

Jinze Bai et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ivana Balažević, Carl Allen, and Timothy M. Hospedales. 2019. Multi-relational poincaré graph embeddings. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Yash Bansal and Adrian Benton. 2021. Is hyperbolic embedding always better than euclidean? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 5082–5097. Association for Computational Linguistics.

Christopher De Sa, Albert Gu, Christopher Ré, and Frederic Sala. 2018. Representation tradeoffs for hyperbolic embeddings. *arXiv preprint arXiv:1804.03329*.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. *arXiv preprint arXiv:1804.01882*.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. 2019. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 3231–3241, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 6338–6347. Curran Associates, Inc.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. *arXiv preprint arXiv:1806.03417*.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. 2018. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4460–4469. PMLR.

Rik Sarkar. 2011. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing: 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21-23, 2011, Revised Selected Papers*, volume 7034 of *Lecture Notes in Computer Science*, pages 355–366. Springer.

Chandni Saxena, Mudit Chaudhary, and Helen Meng. 2022. Cross-lingual word embeddings in hyperbolic space. *arXiv preprint arXiv:2205.01907*.

A. Instruction for Synset Alignment

GPT-5-mini was used as an assistive tool for synset alignment between CWN and OEWN. The alignment process involved only lexical semantic mappings between publicly available lexical resources, with no personal data or human-sensitive information involved.

Given a Chinese sense definition and example sentence, and an English WordNet lookup tool, the model should iteratively search for possible candidate words and synsets, then identify the best match to align the CWN senses with their corresponding PWN synsets. The following snippet defines the system prompt used in the alignment pipeline.

```
Find the best synset in WordNet for the following Chinese word.
First, use tools to look up possible words in the WordNet.
Based on the returned synsets, choose the best match.
You can use tools multiple times if needed, but stop after 5 rounds.
If no good match is found, return "None".

The POS tag "VH" in Chinese Wordnet is a stative verb,
which might map to either "a" (adjective), "v" (verb), or "r" (adverb) in Wordnet.

Along with the synset ID, provide your evaluation of the alignment with a brief
    explanation.
There are three levels of alignment: "exact", "close", and "related".

For example:
    - "cat" (the animal) and 「貓」(一種動物) is counted as "exact".
    - "dumpling" (small boiled/steamed dough, often filled) and 「包子」(一種食物) are
     counted as "close".
    - "egg pancake" (a kind of food made from egg and flour) and 「蛋餅」(一種食物) are
     counted as "related".
    - 「騎樓」 (a type of covered sidewalk) has no good match in Wordnet; return "
    None".

Output a JSON object:
{
  "synset": "00201234-n" or "None",
  "alignment": "exact" | "close" | "related" | "None",
  "explanation": "brief explanation"
}
```

B. Training Hyperparameters

We train Poincaré and Euclidean embeddings using the implementation from Nickel and Kiela (2017), with modifications to support a two-stage learning rate schedule and to save multiple checkpoints during training. Training was performed using Python 3.6 and PyTorch 1.0.0, while all evaluation scripts use Python 3.13 and PyTorch 2.8. All models were trained on a single NVIDIA RTX A5000 GPU, with full training (3000 epochs) requiring approximately 3.5 hours. Table 4 summarizes the key hyperparameters, and the complete training command is provided below for reproducibility.

We employ a two-stage learning rate schedule: 5.0 (0.5 for Euclidean) for epochs 1–1500 and 2.5 (0.25) for epochs 1501–3000. This schedule allows the model to make larger updates during initial training while taking more conservative steps as the embedding space stabilizes. The burn-in period uses a substantially reduced learning rate (multiplier of 0.01) for the first 20 epochs to prevent early training instability.

Hyperparameter	**Value**
Manifold	Poincaré & Euclidean
Dimension	20
Epochs	3000
LR (1–1500)	5.0 & 0.5
LR (1501–3000)	2.5 & 0.25
Batch size	64
Negative samples	50
Burn-in epochs	20
Burn-in LR mult.	0.01
Neg. weight mult.	0.1
Dampening	1.0
Training threads	1
Data proc. threads	8

Table 4: Training hyperparameters.

C. Additional Evaluation Metrics

This appendix presents additional evaluation metrics and visualizations that complement the main results in Section 4.2. All figures show results based on the final 10 training checkpoints (epochs 2775–3000) with error bars or shaded regions representing ± 1 standard deviation.

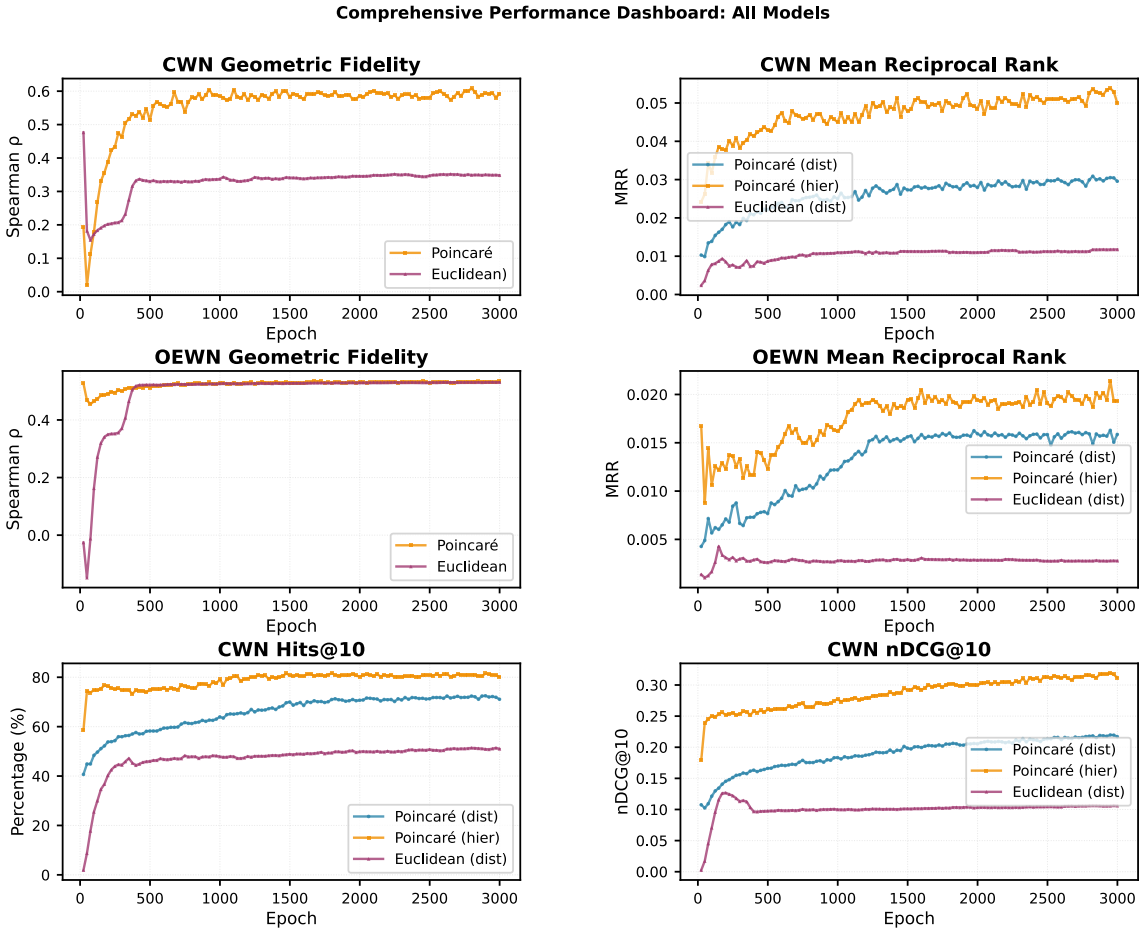


Figure 5: **Comprehensive performance dashboard across all evaluation metrics.** Comprehensive performance dashboard comparing three embedding models across training epochs. The six-panel layout presents: (top row) geometric fidelity (Spearman ρ) and Mean Reciprocal Rank for CWN; (middle row) geometric fidelity and MRR for OEWN; (bottom row) Hits@10 and nDCG@10 for CWN. All panels show epoch-by-epoch evolution for Poincaré embeddings with distance-only ranking (blue), Poincaré embeddings with hierarchical filtering (orange), and Euclidean embeddings with distance-only ranking (magenta). The dashboard demonstrates consistent superiority of Poincaré embeddings across both ranking performance metrics (MRR, Hits@10, nDCG@10) and geometric fidelity measures (Spearman ρ) on both CWN and OEWN test sets.

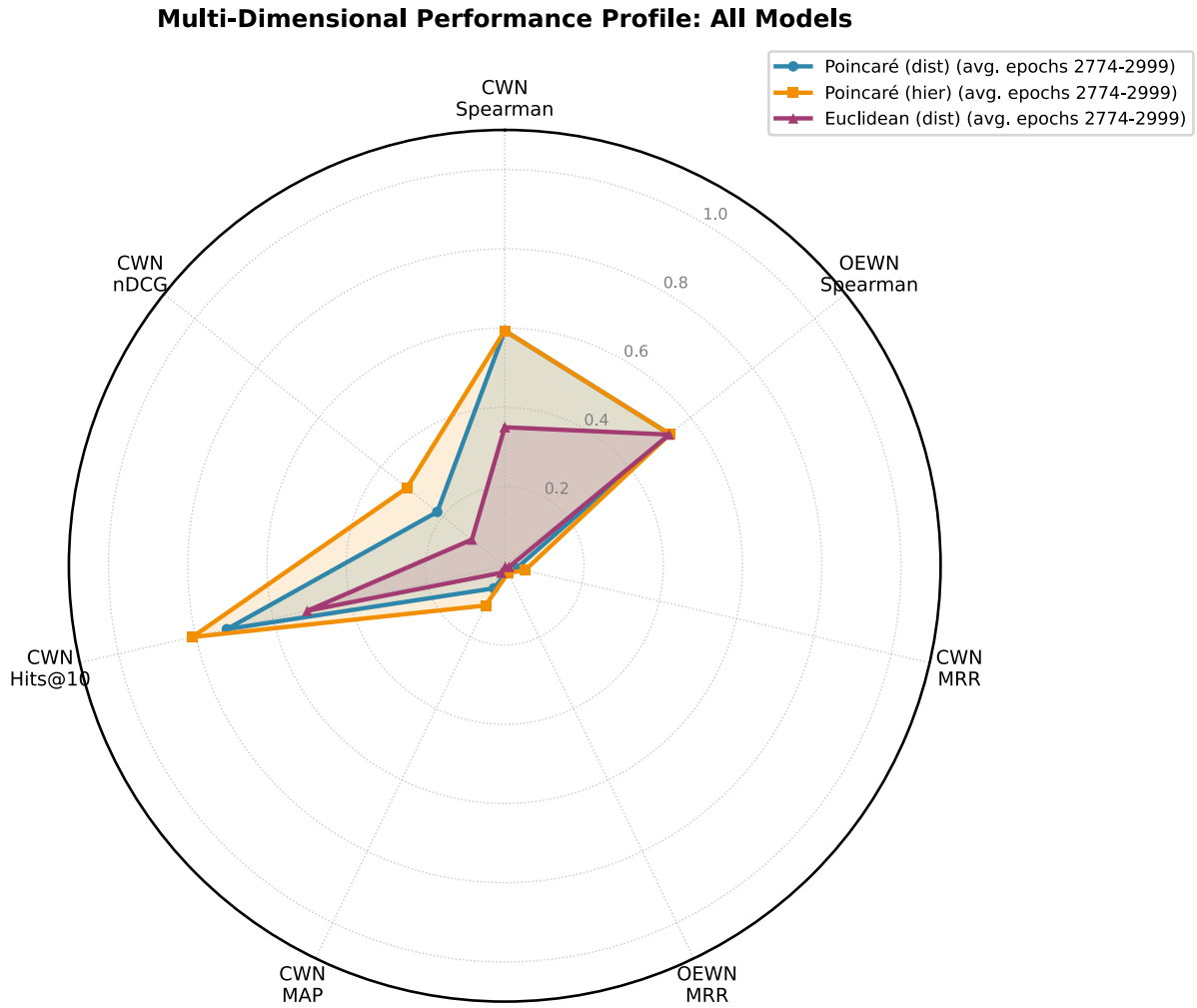
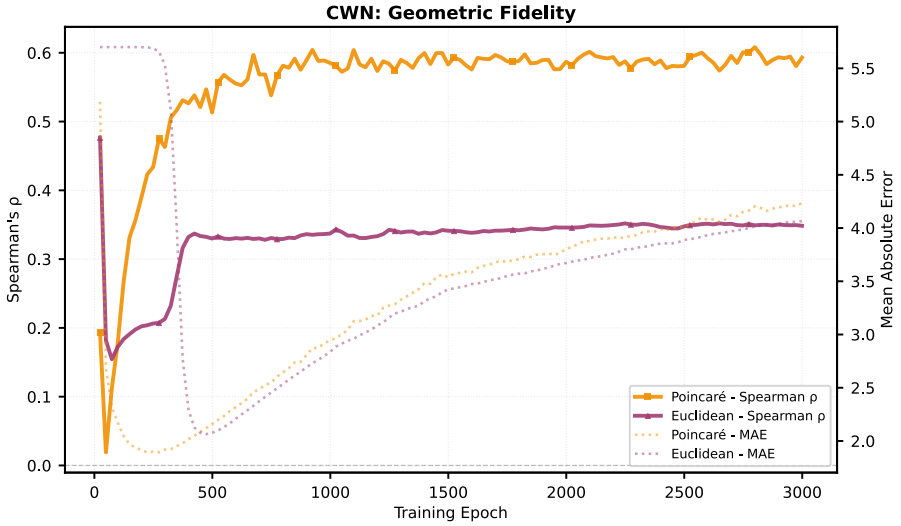
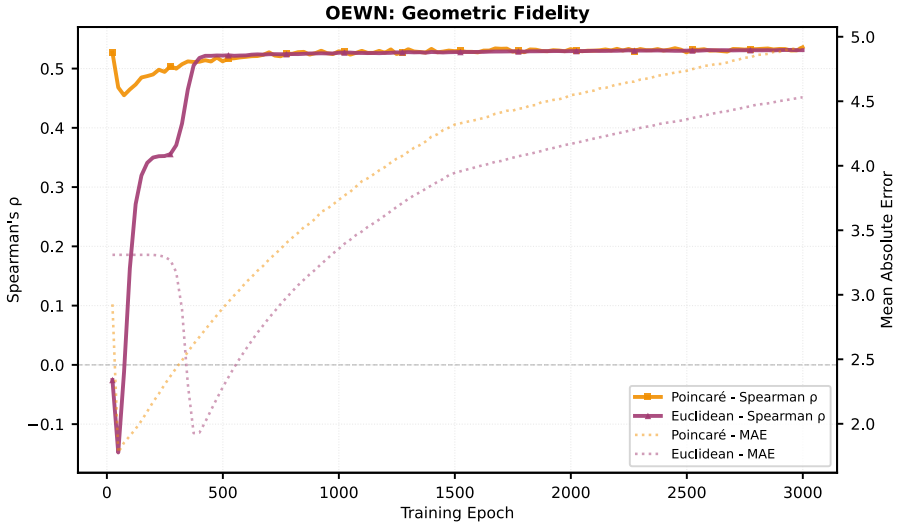


Figure 6: **Multi-dimensional performance profile comparing embedding geometries.** Multi-dimensional performance profile comparing three embedding models across seven key evaluation metrics. The radar chart displays mean performance averaged over the final 10 training checkpoints (epochs 2775–3000) across: CWN geometric fidelity (Spearman ρ), OEWN geometric fidelity, CWN Mean Reciprocal Rank (MRR), OEWN MRR, CWN Mean Average Precision (MAP), CWN Hits@10, and CWN normalized Discounted Cumulative Gain (nDCG@10). Values are averaged to provide stable final performance estimates. Poincaré embeddings with distance-only ranking (blue) and hierarchical filtering (orange) consistently outperform Euclidean embeddings (magenta) across all dimensions, demonstrating the effectiveness of hyperbolic geometry for representing hierarchical semantic structures.



(a) CWN Geometric Fidelity



(b) OEWN Geometric Fidelity

Figure 7: **Geometric Fidelity: Spearman's Correlation (ρ) and Mean Absolute Error (MAE).** Dual-axis comparison of Spearman's rank correlation coefficient (ρ, solid lines, left y-axis) and Mean Absolute Error (MAE, dashed lines, right y-axis) across training epochs. The left y-axis (ρ) measures how well embedding distances preserve hierarchical structure (higher is better). The right y-axis (MAE) measures the average magnitude of distance prediction errors (lower is better). **Panel a:** For Chinese WordNet (CWN), Poincaré embeddings (orange) demonstrate superior correlation with hierarchical relationships ($\rho \approx 0.60$) compared to Euclidean embeddings (magenta, $\rho \approx 0.35$), while both models show competitive MAE performance. The dual-axis visualization reveals that hyperbolic geometry's advantage lies primarily in preserving hierarchical structure rather than minimizing raw distance errors. **Panel b:** For Open English WordNet (OEWN), both Poincaré (orange) and Euclidean (magenta) embeddings converge to similar performance on OEWN's lexical hierarchy ($\rho \approx 0.53$), with Poincaré maintaining slightly higher correlation. The comparable performance on both metrics suggests that OEWN's hierarchy may be more amenable to Euclidean representation compared to CWN, or that the hyperbolic advantage is less pronounced for this particular test set.

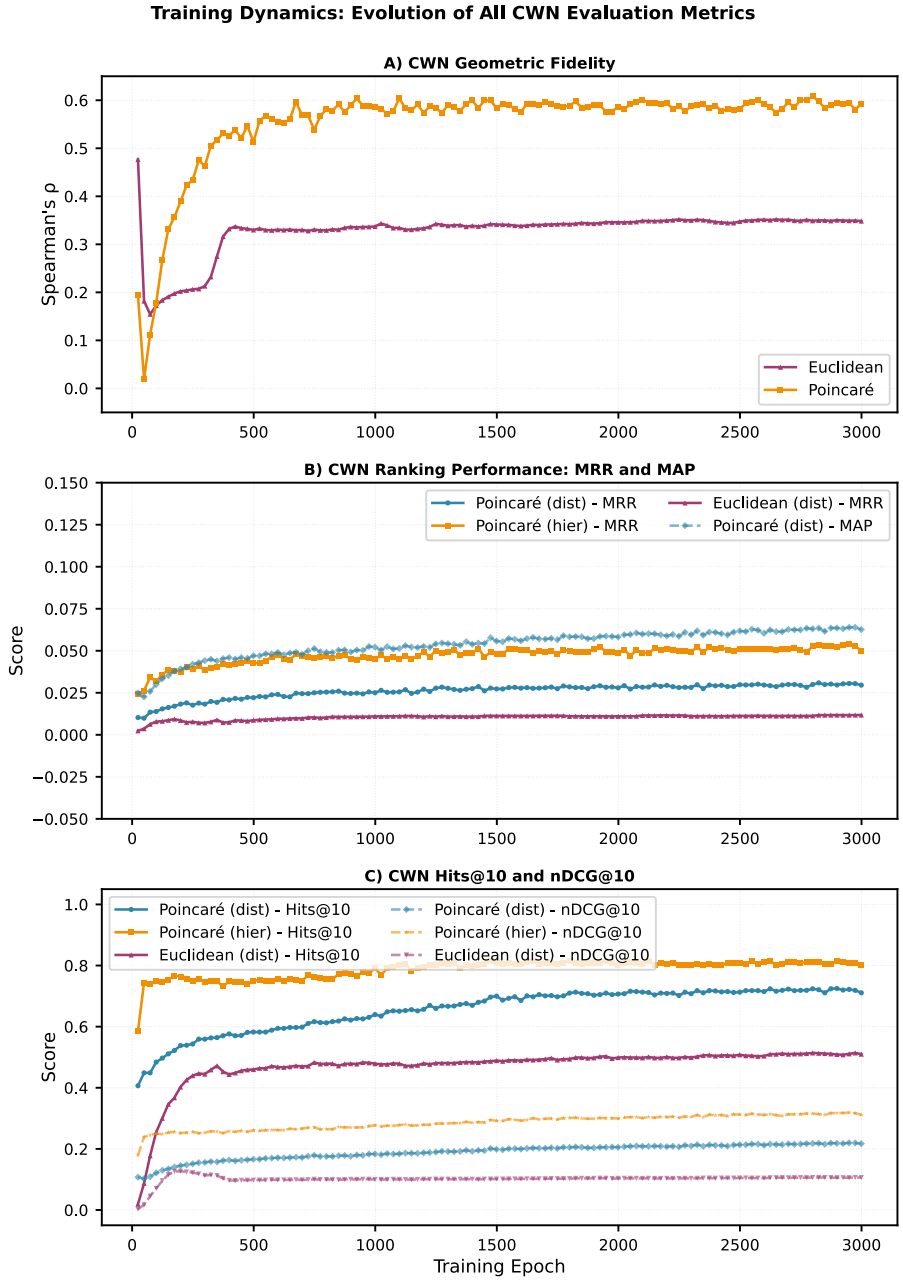
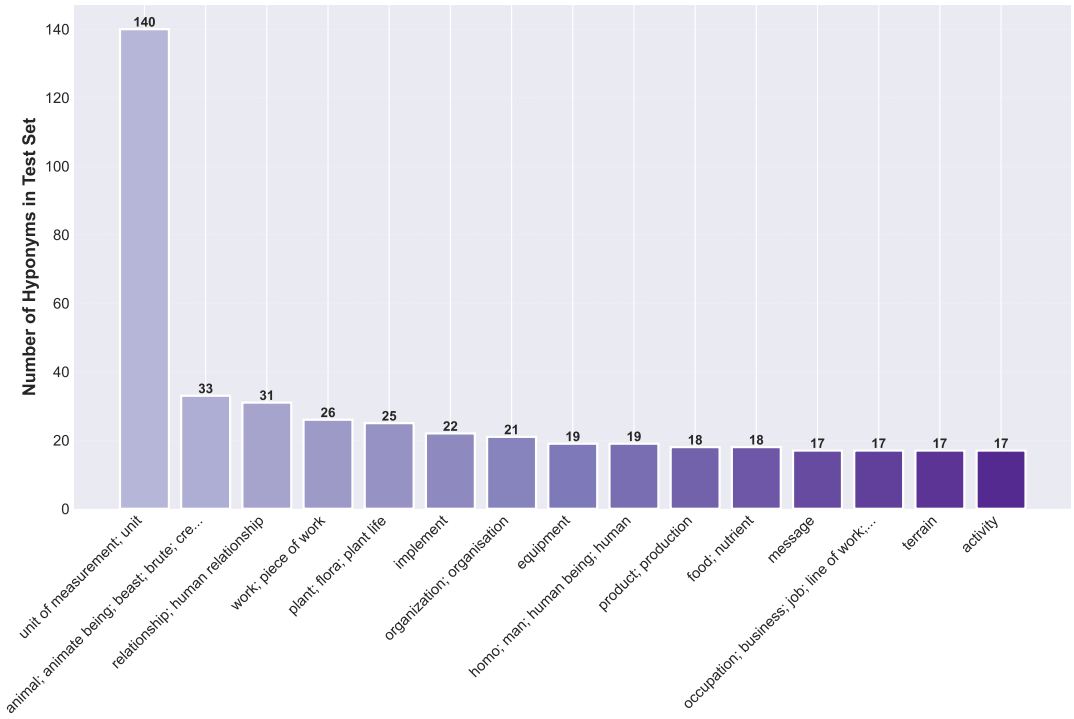
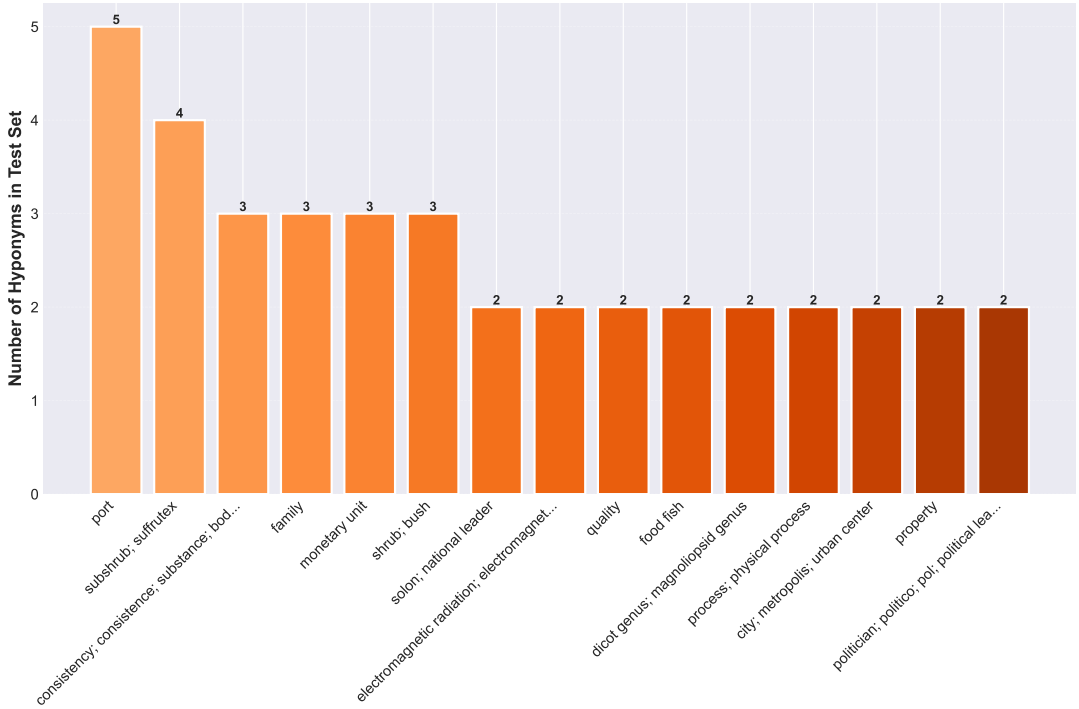


Figure 8: **Training dynamics and convergence behavior across 3,000 epochs.** Evolution of key evaluation metrics across training epochs for three embedding models. Panel (A) displays CWN geometric fidelity measured by Spearman's rank correlation (ρ), showing Poincaré embeddings achieving $\rho \approx 0.60$ compared to Euclidean at $\rho \approx 0.35$. Panel (B) shows ranking quality metrics with Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), where Poincaré with hierarchical filtering achieves the highest MRR (≈ 0.05) and MAP (≈ 0.11). Panel (C) presents retrieval performance with Hits@10 (solid lines) and normalized Discounted Cumulative Gain at rank 10 (nDCG@10, dashed lines). Poincaré embeddings with distance-only ranking (blue) and hierarchical filtering (orange) demonstrate rapid convergence and sustained superior performance compared to Euclidean embeddings (magenta) across all metrics. The visualization reveals that Poincaré models achieve both better geometric fidelity and ranking performance, with hierarchical filtering providing notable improvements particularly in ranking-based metrics, reaching Hits@10 of 80% compared to 50% for Euclidean embeddings.

D. Top 15 Synsets by Number of Hyponyms



(a) CWN Test Set



(b) OEWN Test Set

Figure 9: **Top 15 Synsets by Number of Hyponyms**