Exploring the Small World of Word Embeddings: A Comparative Study on Conceptual Spaces from LLMs of Different Scales

Anonymous ACL submission

Abstract

A conceptual space represents concepts as nodes and semantic relatedness as edges. Word embeddings, combined with a similarity metric, provide an effective approach to constructing such a space. Typically, embeddings are derived from traditional distributed models or encoder-only pretrained models, whose objectives directly capture the meaning of the current token. In contrast, decoder-only models, including large language models (LLMs), predict the next token, making their embeddings less directly tied to the current token's semantics. Moreover, comparative studies on LLMs of different scales remain underexplored. In this paper, we construct a conceptual space using word embeddings from LLMs of varying scales and comparatively analyze their properties. We establish a network based on a linguistic typology-inspired connectivity hypothesis, examine global statistical properties, and compare LLMs of varying scales. Locally, we analyze conceptual pairs, WordNet relations, and a cross-lingual semantic network for qualitative words. Our results indicate that the constructed space exhibits small-world properties, characterized by a high clustering coefficient and short path lengths. Larger LLMs generate more intricate spaces, with longer paths reflecting richer relational structures and connections. Furthermore, the network serves as an efficient bridge for cross-lingual semantic mapping.

1 Introduction

The conceptual space framework, which represents concepts (i.e., words) as nodes and connects them based on their proximity, has been extensively applied in cognitive science, typological linguistics, and related fields (Gärdenfors, 2000, 2014; Nosofsky, 1986, 1987, 1992; Shepard, 1964, 1987; Croft, 2003; Haspelmath, 2003). This framework aims to uncover the alignment between surface-level linguistic structures and deep cognitive processes, offering a spatial visualization of brain activities or concept representations.

Existing approaches employ both manual and automated methods to define nodes and edges, thereby constructing the conceptual space. In manual methods, linguists either establish direct connections between nodes based on connectivity hypotheses (Croft, 2001) or determine similarity by computing the co-occurrence frequency of two concepts (Cysouw, 2007), subsequently generating a low-dimensional network using techniques such as Principal Component Analysis (PCA) (Abdi and Williams, 2010) or Multidimensional Scaling (MDS) (Goldstone, 1994). However, these manual approaches are labor-intensive and inadequate for constructing large-scale conceptual spaces.

To address the aforementioned limitations, automated methods utilize embedding models to improve efficiency. Traditional embedding models, such as static word2vec (Church, 2017), represent words as dense vectors. However, these models rely on limited datasets and often fail to capture nuanced word meanings. In contrast, pretrained models like BERT (Devlin et al., 2019) have demonstrated greater effectiveness in word representation (Devlin et al., 2019; Tenney et al., 2019) and have become widely adopted across related fields (Moullec and Douven, 2025). As a purely encoder-based model, BERT focuses on recovering masked words, which gives it a distinct advantage in constructing conceptual spaces.

Recently, Large Language Models (LLMs) have demonstrated remarkable performance across a variety of understanding and generation tasks (OpenAI, 2023). However, research on word embeddings within LLMs remains relatively limited. Typically, LLMs are decoder-only models trained with the objective of next-token prediction. Consequently, it remains unclear to what extent the embeddings capture the meaning of the current token, as opposed to simply transferring information to the next token (Liu et al., 2024a). Furthermore,

despite their success, LLMs still face challenges in terms of interpretability (Zou et al., 2023). In this context, we aim to offer a cognitive perspective for interpreting LLMs comparatively by constructing a conceptual space based on initial input representations from LLMs of different scales.

In this paper, we construct a conceptual space based on LLM input embeddings and analyze its properties comparatively from both global and local perspectives. Specifically, we treat the LLM vocabulary as a set of concepts, with the input embeddings serving as node representations. Using similarity metrics, we first build a complete graph, which we then sparsify based on the minimal connectivity hypothesis inspired by semantic map models (Haspelmath, 2003; Croft, 2003). We compare the conceptual spaces of LLMs with similar architectures but different parameter scales by calculating global network statistics. Our findings show that LLM embeddings form a small-world network with high clustering coefficients and low average path lengths. LLMs with larger parameter scales tend to have more complex structures, with longer paths and richer relationships. To assess the practical utility of the conceptual space, we extract local subgraphs for scenarios such as: (1) common concepts, (2) WordNet relations, and (3) a cross-lingual case study on qualitative words. The consistency of these subgraphs with human annotations confirms the effectiveness of our constructed conceptual space. In conclusion, our contributions are as follows:

- We propose a comparative study on conceptual spaces constructed from LLMs of different scales based on the proposed connectivity hypothesis.

- We design three distinct scenarios and conduct an extensive evaluation of the conceptual space, considering both global and local perspectives, while comparing models of different scales.

- We demonstrate that the conceptual spaces align with human perception and provide an effective representation of concepts and their relationships.

2 Related Work

2.1 Conceptual Space Modeling

Conceptual space modeling has been extensively studied and applied in fields such as cognitive sci-

ence (Gärdenfors, 2000, 2014; Nosofsky, 1986, 1992), linguistic typology (Croft, 2001; Haspelmath, 2003), and neuroscience (Caglar, 2021). One widely used framework, the Conceptual Space Framework (CSF) proposed by Gärdenfors (2000, 2014), introduces the basic concepts of similarity space and conceptual space, with the latter being a prototypical realization based on the former. Subsequent work has focused on constructing similarity spaces, including the representation of instances and the distance metrics among them. Multidimensional scaling (Borg and Groenen, 1999) and spatial arrangement methods (Goldstone, 1994) are two common approaches based on pairwise similarity judgments for a set of items. However, these methods are often manual and incur high costs due to the need for extensive data collection or the cognitive demands they impose. Alternatively, a more efficient approach leverages language models and word embeddings, such as word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017), BERT (Devlin et al., 2019), and even LLMs (Touvron et al., 2023). Notably, embeddings directly from LLMs exhibit weaker correlation with human-annotated datasets compared to prompt-guided constructions. Our paper also focuses on embeddings from LLMs, but with more systematic evaluations and a broader set of concepts [1].

2.2 Semantic Map Models

Semantic map modeling is another framework for constructing conceptual spaces based on cross-lingual co-occurrence of concepts in linguistic forms. These forms can include content words (Guo, 2012b; Cysouw, 2007; Perrin, 2010), function words (Zhang, 2017), or constructions (Malchukov et al., 2007). The concepts are typically represented by the grammatical (Zhang, 2017) or content (Guo, 2012b) meanings of these forms. The conceptual space can be constructed in either a bottom-up or top-down manner. A classical bottom-up approach is based on the connectivity hypothesis (Croft, 2001; Haspelmath, 2003; Teng, 2015), which posits that concepts involved in a single linguistic form should be connected within the corresponding subgraph. The overall space is then built incrementally, edge by edge. Alternatively, a top-down approach (Liu et al., 2024b) first

---

[1] In this paper, the term "concept" refers to a word without context, similar to the instances in CSF. Thus, our conceptual spaces are more akin to similarity spaces within the CSF framework.

constructs a similarity graph based on the strength of co-occurrence, and then sparsifies the graph according to a refined connectivity constraint. In this paper, we treat a concept as a single token without explicitly matching form-meaning pairs, and we adopt an efficient top-down approach with slightly modified constraints to build the conceptual space.

2.3 Word Embedding and Representation

Contemporary language models represent words or subtokens using continuous vectors based on distributed semantics (Boleda, 2020). These high-dimensional vectors can be static (Mikolov et al., 2013; Bojanowski et al., 2017) or context-sensitive (Devlin et al., 2019). While effective, they lack the interpretability of linguistic feature-based representations (Petersen and Potts, 2023). Word embeddings exhibit elegant linear relationships (Mikolov et al., 2013), high similarity with human judgments (Vulić et al., 2020), and meaningful representations (Turney and Pantel, 2010). Static embeddings are particularly suitable for offline, context-free words, especially monosemous ones. These embeddings are used in the input and output layers of LLMs. Previous research has explored their linear properties (Han et al., 2024), conceptual space construction (Moullec and Douven, 2025), and other aspects. In this paper, we focus on LLM input embeddings, offering a more diverse and systematic evaluation through the lens of conceptual spaces.

3 Approach

In this section, we first introduce the basic notions and construct a complete graph connecting all conceptual nodes. We then sparsify the graph based on the revised connectivity hypothesis. Finally, we define global and local metrics to evaluate the conceptual space. The overall pipeline is illustrated in Figure 1.

3.1 Basic Notions

We define a conceptual space $\mathcal{G} = \{V, E\}$, where V and E are sets of nodes and edges, respectively. Each node $v \in V$ represents a concept, which can be realized by a token, word, or sense. Each edge $e(u, v) \in E$ connects a pair of nodes (u, v), reflecting their degree of association (Guo, 2012a). If a path $p(u, v)$ exists between nodes u and v, they are connected, with path length L defined as the number of edges along the path. If no path

exists, $L = \infty$. A conceptual space is considered connected if every pair of nodes is connected.

A subgraph $\mathcal{G}' = \{V', E'\}$, where $V' \subset V$ and $E' \subset E$, reflects the local topology of $\mathcal{G}$ and typically represents a specific semantic domain, such as adverbs (Zhang, 2017), color adjectives (Gärdenfors, 2014), or qualitative words (Perrin, 2010). Similarly, a subgraph is connected if every pair of nodes has a path.

We define a metric M on $\mathcal{G}$ to measure the association or similarity between nodes. A common metric is cosine similarity [2], widely applied in similarity-related tasks.

3.2 Complete Graph

We use an LLM to extract input embeddings $\mathcal{E}$ for all tokens in its vocabulary $\mathcal{V}$, treating each token (the minimal computational unit) as an individual concept. After obtaining the vectorized embeddings, we compute the cosine similarity between every pair of nodes to define edge weights. Additionally, we apply centering by subtracting the average vector from each embedding to address anisotropy (Ethayarajh, 2019). This results in a complete graph $\mathcal{C}$, where every pair of concepts is connected.

3.3 Conceptual Space

We derive a sparsified graph, denoted as the conceptual space $\mathcal{G}$, from the complete graph $\mathcal{C}$. We propose a minimum connectivity hypothesis, which states that $\mathcal{G}$ must remain connected while using the fewest edges possible. The "connectivity" ensures that every pair of concepts is connected, forming a valid space. The "minimum" condition favors sparse connections, inspired by the top-down construction of semantic map models, which even use trees (with the least number of edges) to maintain connectivity. To achieve this sparsity, we rank edges by weight and retain the top K ratio of edges, as higher weights indicate more important connections.

A well-defined $\mathcal{G}$ is also a discrete topological space $\{\mathcal{G}, \mathcal{T}\}$, where $\mathcal{T}$ is the collection of all subsets. We define a subgraph $\mathcal{G}'$ as a subset of $\mathcal{G}$, and it is considered an open set. This is because the intersection and union of any two subgraphs $\mathcal{G}_A$ and $\mathcal{G}_B$ still belong to $\mathcal{T}$:

$$\forall \mathcal{G}_A, \mathcal{G}_B \in \mathcal{T}, \quad \mathcal{G}_A \cap \mathcal{G}_B \in \mathcal{T}, \quad \mathcal{G}_A \cup \mathcal{G}_B \in \mathcal{T}. \quad (1)$$

---

[2]While cosine distance violates the triangle inequality required by strict distances, we relax this constraint due to its simplicity and widespread use.
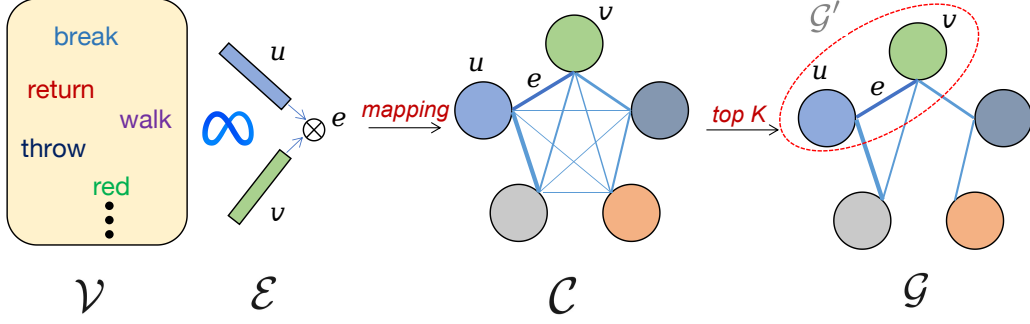
Figure 1: Outline of our conceptual space construction. First, we extract the input word embeddings ($\mathcal{E}$) for the LLM vocabulary ($\mathcal{V}$). Next, we build a complete graph $\mathcal{C}$ by calculating the cosine similarity between all embedding pairs. Finally, we retain edges based on similarity, from highest to lowest, until the graph $\mathcal{G}$ is connected. We then focus on specific connected subgraphs $\mathcal{G}'$ representing certain domains at a local level.

This is ensured by the "connectivity" condition, while a "minimum" topological space is required for the conceptual structure.

3.4 Evaluation

We evaluate the conceptual space from both global and local perspectives.

Globally, we compute network statistics to analyze basic properties, connectivity, and small-world characteristics of the spaces built by two models. Small-world characteristics are indicated by a higher clustering coefficient and a shorter shortest path, which are described in detail in Section 5.1.

Locally, we analyze a subgraph $\mathcal{G}'$ of the conceptual space $\mathcal{G}$ in three scenarios. Scenario 1 examines common concepts across ten semantic categories, each containing monosemous words, comparing shortest paths within and between groups. Scenario 2 explores shortest-path connections for various WordNet relations. Scenario 3 evaluates a conceptual space of qualitative words, comparing it to the corresponding LLM subgraph. Beyond topology and connectivity, we assess node degree correlations and measure recall and precision against the ground truth.

4 Experimental Design

4.1 Large Language Models

We adopt the Llama series as our LLMs, including Llama2-7B and Llama2-70B (Touvron et al., 2023). The dimension of the input embedding is 4096 and 8192, for Llama2-7B and Llama2-70B respectively. Also, they share the vocabulary for both models, with the size of vocabulary 32,000. The tokens in the vocabulary are obtained by Byte Pair Encoding (Sennrich et al., 2016), merging the frequent characters. Thus, many tokens are part of a whole word. Besides, tokens with a whitespace or appearing at the beginning of a sentence are different from those in other places, i.e., the end part of a word. For example, "man" in "policeman" and "man" are different units in the vocabulary. We identify the token appearing the end part of a word by add "#" at the beginning of the token, such as "#man".

4.2 Scenario 1: Common Concepts

We collect nine semantic groups to represent common concepts, each containing 10 tokens or concepts: NUMBER, NAME, MONTH, COLOR, CITY, NATION, PLACE, HUMAN, and FURNITURE. Additionally, we include a semantic group of RANDOM concepts. A full list of the concepts is provided in Table 4 in Appendix A.1. This scenario primarily examines the length of the shortest path within and between semantic groups.

4.3 Scenario 2: WordNet Relations

In Scenario 2, we explore a subgraph of WordNet instances and their structural relationships. We extract a subset from the public WN18 dataset (Bordes et al., 2013), consisting of 40,943 WordNet synsets and 18 relation types. The dataset is filtered by the following conditions: (1) after converting synsets to words, the words must be in the Llama vocabulary; (2) only the first sense of each synset is retained for its stereotypical meaning; (3) a relation must involve at least 10 words; (4) symmetric relation pairs, such as hypernym and hyponym, are merged. Additionally, we define two wordform relations: *tokenization variant*, distinguishing tokens with and without a leading underscore (e.g., "man" vs. "#man"), and *uppercase variant*, differentiating

capitalized and non-capitalized forms (e.g., "red" vs. "Red"). In total, we consider eight relation types, listed in Table 1.

Index	Relation Type	Count
A	Member of Domain Topic	51
B	Verb Group	10
C	Hypernym	464
D	Has Part	22
E	Also See	95
F	Derivationally Related Form	388
G	Tokenization Variant	1685
H	Uppercase Variant	1788

Table 1: WordNet Relations and Instance Counts

4.4 Scenario 3: SMM of Qualitative Words

In Scenario 3, we utilize a cross-lingual semantic map for adjectives and qualitative words (Perrin, 2010), which includes 22 African languages, French, and English. Polysemes in each language are connected to indicate conceptual proximity. The domain spans dimension, age, value, color, etc., with capitalized English words representing concepts, such as BIG, SMALL, LONG, SHORT, WIDE, and DEEP for dimension. We filter out words not present in the Llama vocabulary, resulting in 75 concepts. The final set of concepts and the human-annotated graph are presented in Appendix A.2.

5 Results and Analysis

In this section, we first construct the conceptual space based on the minimum connectivity approach described in Section 3.3. We then evaluate the space in three distinct scenarios.

5.1 Graph Construction

Choice of K. To ensure the graph is minimally connected, we extract the top K ratio of edges, where the edge weights are determined by cosine similarity. We incrementally increase the value of K while monitoring the number of connected components (CC), as shown in Figure 2. The graph first becomes connected when the log of the number of CC reaches zero. In our experiments, we selected $K = 0.002$, at which point both models become connected. This choice ensures that the same number of edges are used for both models, making the comparison fair.
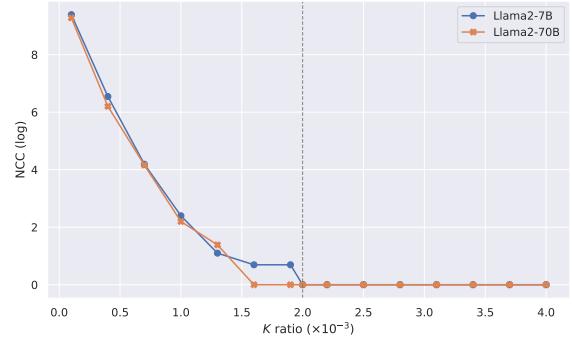

Figure 2: Logarithm of the number of connected components as the top K ratio increases for Llama2-7B and Llama2-70B. A value of zero indicates a fully connected network, while the dotted line marks the first ratio at which both models become connected.

Statistics	Llama2-7B	Llama2-70B
Basic		
#Nodes	32,000	32,000
#Edges	1,024,000	1,024,000
Avg. Degree	64	64
Std. Degree	68.39	58.96
Weighted		
Avg. Degree_W	8.76	12.23
Std. Degree_W	13.78	14.02
Threshold	0.095	0.147
Small-world		
GCC ($\uparrow$)	0.325	0.215
ALCC ($\uparrow$)	0.183	0.174
Diameter ($\downarrow$)	6	6
ASPL ($\downarrow$)	3.392	3.353

Table 2: Statistics for Llama2-7B and Llama2-70B. In the "Small-world" section, GCC refers to global clustering coefficient, ALCC to average local clustering coefficient, and Diameter and ASPL refer to the longest and average shortest path lengths, respectively.

Global Statistics. We present the statistics of the conceptual spaces for both models in Table 2, divided into three parts. The **Basic** section includes the number of nodes (#Nodes), edges (#Edges), and the average (Avg. Degree) and standard deviation (Std. Degree) of degrees. The **Weighted** section calculates the average and standard deviation of weighted degrees (also called Strength or Traffic), along with the equivalent threshold—the minimum weight value in the final graph G. The last section focuses on **Small-world** effects. The
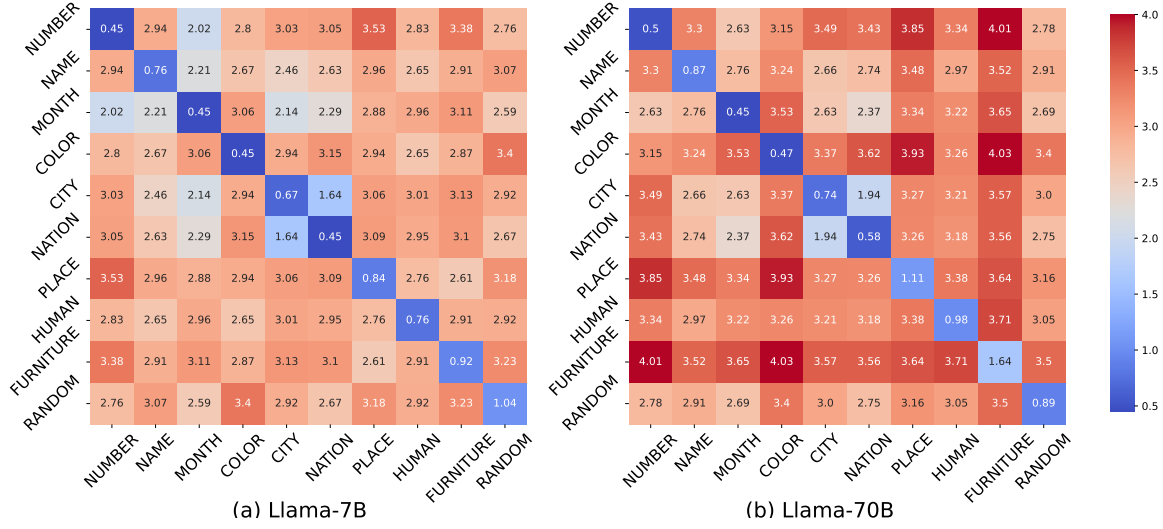
Figure 3: Shortest path lengths among semantic groups for Llama2-7B (left) and Llama2-70B (right).

global clustering coefficient (GCC) measures graph transitivity—the fraction of actual triangles among all possible triangles in G. The local clustering coefficient (ALCC) is calculated as the average of actual connections within neighbors for all nodes. The network diameter is the longest path between any two nodes. Smaller diameters and larger GCC/ALCC values indicate stronger small-world effects. We also calculate the average shortest path length (ASPL) for both models.

Llama2-7B and Llama2-70B share similar graph structures, both using the same number of edges. The average degree is 64, but 7B has a flatter degree distribution with a larger standard deviation. In the weighted version, 7B's degree distribution is more concentrated. As shown in Figure 11 (Appendix A.3), 7B exhibits a long-tailed distribution, indicating fewer high-degree (central) nodes.

Both models exhibit strong small-world clustering, with high GCC and ALCC values and a short diameter. In contrast, random networks with the same edge count have much lower GCC (0.0032) and ALCC (0.0020). The observed diameter of 6 aligns with the Six Degrees of Separation theory (Milgram, 1967), commonly found in social networks (Watts and Strogatz, 1998) and web structures (Réka Albert and Barabási, 1999).

5.2 Scenario 1

In this scenario, we evaluate the conceptual spaces of both models by calculating the shortest path length between any pair of nodes within and between semantic groups. The shortest path is computed using Dijkstra's algorithm from the Net-
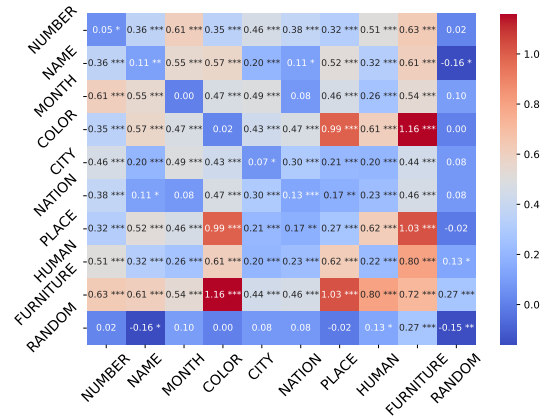

Figure 4: Shortest Path Length Difference (Llama2-70B Minus Llama2-7B) Across Semantic Groups. The number of stars indicate the degree of the significance level of the difference.

workX package [3]. The average lengths for Llama2-7B and Llama2-70B are shown in Figure 3. Figure 4 presents the difference heatmap, where the difference is computed as the average length of Llama2-70B minus that of Llama2-7B. The significance of the differences is indicated by the number of stars (one, two, or three), corresponding to p-values of 0.05, 0.01, and 0.001 in the t-test, respectively.

Our results show that Llama2-70B generally has longer path lengths than Llama2-7B, both within and between semantic groups, except in the RANDOM vs. NAME or PLACE comparison. This suggests that Llama2-70B follows more complex

---

[3] https://networkx.org/documentation/stable/reference/algorithms/shortest_paths.html

Figure 5: Paths between "bar" and "library" for Llama2-7B. Edge width reflects weight values. Nodes and edges along the shortest path in terms of the summed weights are marked in red.

Figure 7: Paths between "collection" and "set" in Llama2-7B. The red line represents the shortest path in terms of the summed weights.

Figure 6: Paths between "bar" and "library" for Llama2-70B. Edge width represents weight values. Nodes and edges in the shortest path are highlighted in red. The paths illustrate a complex relationship involving similar word forms, multilingual links, and disambiguation groups.

Figure 8: Paths between "collection" and "set" in Llama2-70B. The red line indicates the shortest path, which follows a more logical transition from a singular concept to a plural one.

paths to uncover relationships. Figures 5 and 6 illustrate this with the six shortest paths from "bar" to "library." The shortest path (red) in Llama2-70B shows greater variation in word form and conjugation. Additionally, its other paths include multilingual instances and diverse connections, aiding disambiguation—for example, distinguishing "library" as a building from its software-related meaning.

When analyzing per group, both models show shorter paths within the same group, and the trends are similar for different group pairs. Among the groups, COLOR shows the shortest paths, while FURNITURE exhibits the longest. This may be due to polysemy in FURNITURE, where terms like "chair" refer to both furniture and a human (e.g., a person referred to as a "chair"). This is further supported by the longer paths when comparing FURNITURE to other groups. Conversely, NATION and CITY tend to have shorter paths.

5.3 Scenario 2

In the second scenario, we evaluate the network for word pairs from different relation types, as shown in Table 1. For each pair, we compute the shortest path and display the length for both models. The significance of the differences is indicated by the number of stars. The results are shown in Figure 9.

The results are similar to Scenario 1: Llama2-70B generally has longer paths than Llama2-7B. For most relations, the length of 70B is significantly longer, except for those related to word form,

7
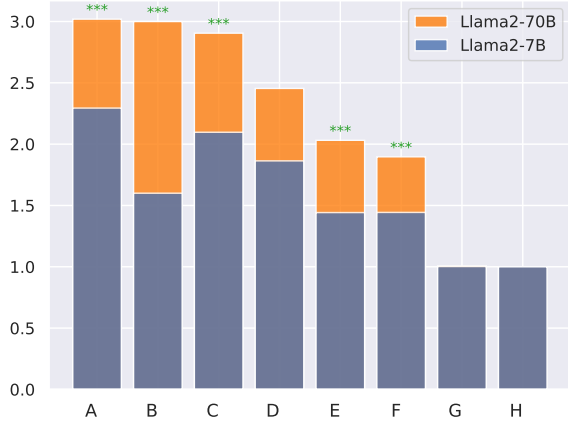
Figure 9: Averaged shortest path length across relation types for both models. The number of stars indicates the significance of the difference.

Statistics	7B	70B	GT
Basic			
#Nodes	75	75	75
#Edges	293	130	37
Avg. Degree	7.813	3.467	0.987
Std. Degree	5.724	3.021	0.959
Weighted			
Avg. Degree_W	0.963	0.611	-
Std. Degree_W	0.736	0.546	-
Avg. Weight	0.123	0.176	-
Connectivity			
#Component (↓)	8	17	39
#Single (↓)	7	16	26
Reference with GT			
Correlation (↑)	0.466	0.449	1
Recall (↑)	0.568	0.378	-
Precision (↑)	0.072	0.108	-

Table 3: Statistics for Llama2-7B, Llama2-70B, and GT across four dimensions. "#Component" represents the number of connected components, and "#Single" indicates the number of nodes with zero degree. In the bottom section, we report degree correlation, recall, and precision.

where both models have direct edges for the word pair. Longer paths appear in *member of domain topic*, *verb group*, and *hypernym* relations, indicating a more complex connection chain. For example, in the *hypernym* relation between "set" and "collection", Llama2-70B shows a more intricate connection, gradually linking the singular "set" to the plural "collection" with some conjugation correlation, as seen in Figures 7 and 8.

5.4 Scenario 3

In Scenario 3, we construct a conceptual space for qualitative words, with a reference (GT) from cross-lingual research. The statistics for the spaces built by Llama2-7B (7B), Llama2-70B (70B), and GT are shown in Table 3.

In general, the spaces constructed by the models have more edges than those built by human experts, with fewer connected components and isolated nodes. This is because experts create the graph by collecting a corpus and adding an edge only when at least three languages exhibit concept co-occurrence. This process is especially challenging for low-resource languages. In contrast, the vectorized concepts from models generate a much denser graph. Furthermore, Llama2-70B tends to be sparser than Llama2-7B. Compared to GT, the automatic space aligns well, showing moderate correlation and coverage (indicated by recall). However, the precision is lower due to the larger number of edges, suggesting that embeddings could be used to initially construct a space, which linguists could later refine.

6 Conclusion

This paper investigates the construction of conceptual spaces using input embeddings from large language models (LLMs). We analyze and compare the network properties of two LLMs with different scales across three scenarios. Our findings show that conceptual spaces can be effectively constructed from embeddings, which exhibit a small-world clustering effect. Additionally, models with more parameters tend to explore longer and more complex paths between concepts, partially supporting the "scaling law" (Kaplan et al., 2020). This study also provides an efficient approach to constructing conceptual spaces, potentially benefiting fields such as language typology and cognitive science.

7 Limitations

We acknowledge several limitations in our work. First, we represent each concept as a word without considering context. However, words can be ambiguous, particularly for homonyms that encom-

pass multiple unrelated meanings. Second, our evaluation is limited to two models, Llama2-7B and Llama2-70B. Results may differ with models of different architectures or parameter scales. Additionally, we focus only on input embeddings and do not explore the properties of output embeddings, which may also capture individual word representations. Finally, the length of the shortest path in conceptual spaces is not a definitive metric for embedding quality, and we plan to explore more sophisticated metrics to better reflect the characteristics of these spaces.

8 Ethics Statement

We do not foresee immediate ethical concerns arising from our research. However, there may be unintended biases in the connections between concepts, such as those involving gender and job ranks. These biases may stem from biased embeddings (Bordes et al., 2013; Bolukbasi et al., 2016).

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2787–2795.

Ingwer Borg and Patrick J. F. Groenen. 1999. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40:277–280.

Leyla Roksan Caglar. 2021. *Conceptual Spaces in the Brain: An Exploration of Structure, Shape, and Organization*. Ph.D. thesis, Rutgers The State University of New Jersey, Graduate School-Newark.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

William Croft. 2003. *Typology and Universals*, second edition edition. Cambridge University Press, Cambridge.

Michael Cysouw. 2007. Building semantic maps: the case of person marking. In Matti Miestamo and Bernhard Walchli, editors, *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, pages 225–248. Mouton, Berlin.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186. Association for Computational Linguistics.

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

R. Goldstone. 1994. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26:381–386.

Rui Guo. 2012a. The minimum connection principles and connection degrees of concepts in semantic maps. In *Proceedings of the Conference on Multilingual and Multifunctional Grammatical Forms*, Peking University.

Rui Guo. 2012b. The typology of adjectives and the grammatical status of chinese adjectives. *Chinese Language Learning*, (5).

P. Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought*. MIT Press.

P. Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The new psychology of language*, volume 2, pages 211–243. Lawrence Erlbaum, Mahwah, NJ.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024a. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024b. A top-down graph-based tool for modeling classical semantic maps: A crosslinguistic case study of supplementary adverbs. *arXiv preprint arXiv:2412.01423*.

Andrej Malchukov, Martin Haspelmath, and Bernard Comrie. 2007. Ditransitive constructions: A typological overview. In *Paper for the Conference on Ditransitive Constructions*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119.

Stanley Milgram. 1967. The small world problem. *Psychology Today*, 2:60–67.

M. Moullec and I. Douven. 2025. Cheaper spaces. *Minds & Machines*, 35(6).

R. M. Nosofsky. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

R. M. Nosofsky. 1987. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13:87–108.

R. M. Nosofsky. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43:25–53.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Loïc-Michel Perrin. 2010. Polysemous qualities and universal networks, invariance and diversity. *Linguistic Discovery*, 8:1–22.

Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of english "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511.

Hawoong Jeong Réka Albert and Albert-László Barabási. 1999. Internet: Diameter of the world-wide web. *Nature*, 401:130–131.

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

R. N. Shepard. 1964. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1:54–87.

R. N. Shepard. 1987. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Ma Teng. 2015. Computer-aided construction of the semantic map model. In Li Xiaofan, Zhang Min, Guo Rui, and et al., editors, *Research on Semantic Maps of Multifunctional Grammatical Forms in Chinese*, pages 194–206. The Commercial Press, Beijing.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Ying Zhang. 2017. Semantic map approach to universals of conceptual correlations: a study on multifunctional repetitive grams. *Lingua Sinica*, 3(1):7.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Common Concepts in Scenario 1

Table 4 lists specific concepts from different semantic groups, with each group containing ten common concepts.

A.2 Conceptual Spaces in Scenario 3

Scenario 3 presents a human-annotated conceptual space for qualitative words, as shown in Figure 10. Each concept is represented by an English word. Nodes are connected if a pair of concepts co-occur as a polysemous word in at least three languages. Nodes marked in red represent federative words, indicating a shared concept with a higher degree.

A.3 Degree Distribution

We show the distribution of node degrees for spaces generated by two models, Llama2-7B and Llama2-70B. Figure 11(a) displays the unweighted degree distribution, while (b) shows the weighted distribution. The results indicate that the 70B model has a less pronounced long-tail distribution, with more nodes having relatively larger degrees.
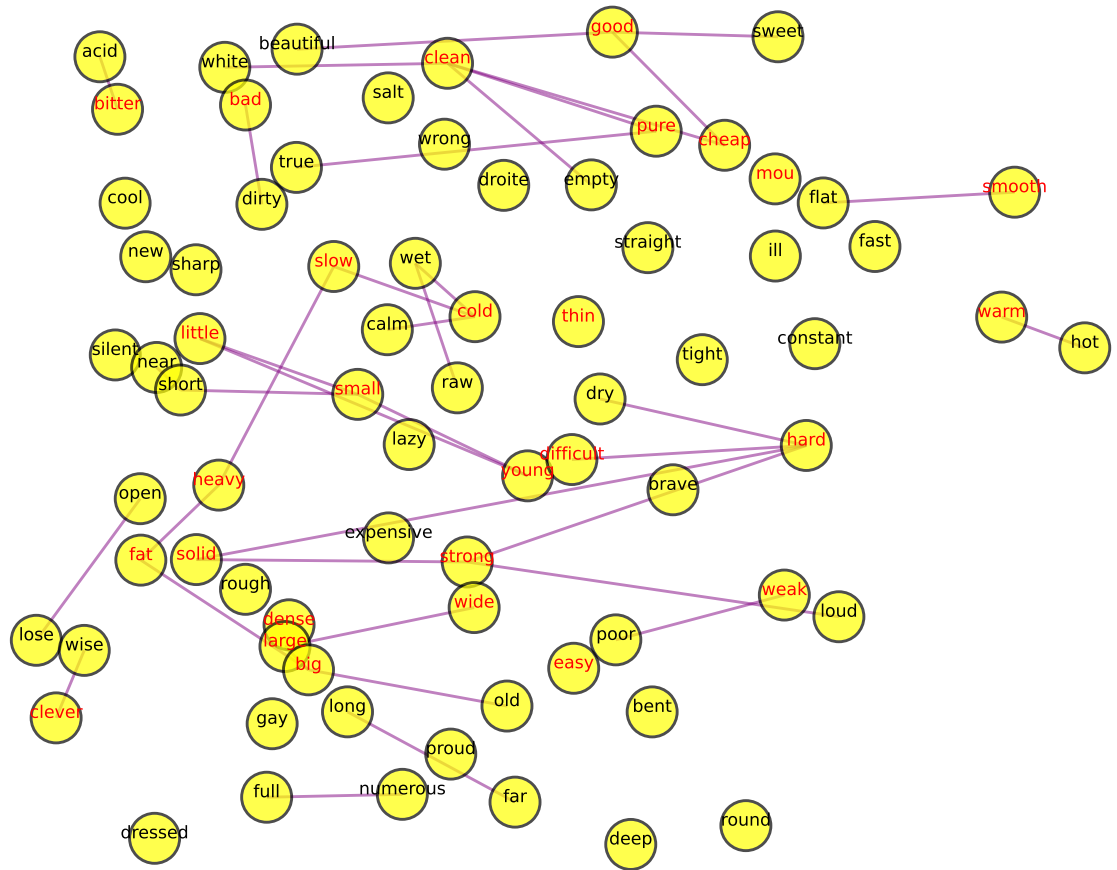
Figure 10: A semantic map for the domain of qualitative words, with federative notions which have a higher degree highlighted in red.
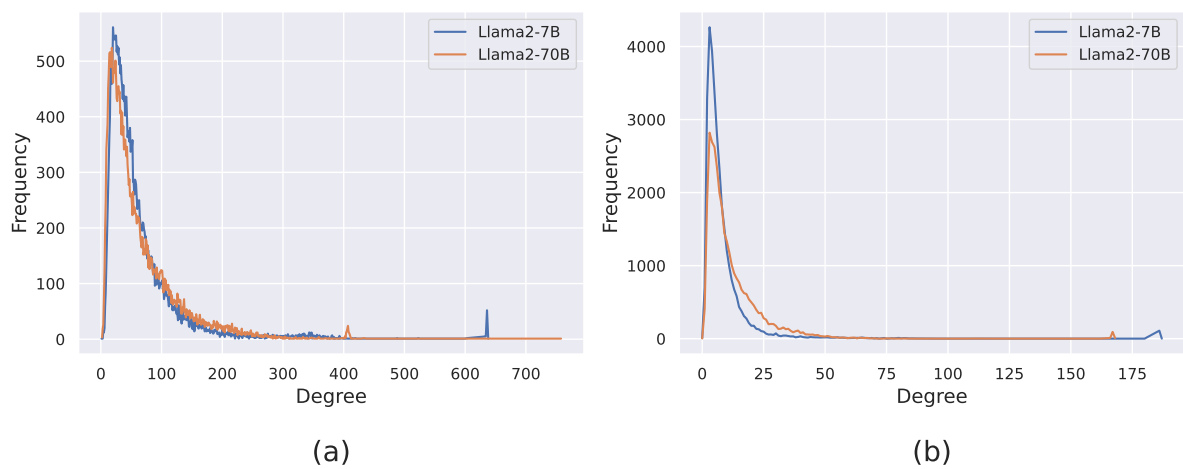

Figure 11: Comparison of the degree distribution for both models: (a) unweighted degree and (b) weighted degree.

Group	Concepts
NUMBER	one, two, three, four, five, six, seven, eight, nine, ten
NAME	Alice, Bob, Carol, Dave, Francis, Grace, Hans, Ivan, Zach, Mike
MONTH	January, February, March, April, May, June, July, August, September, October
COLOR	red, orange, yellow, green, blue, brown, black, white, grey, gray
CITY	Taiwan, York, Cambridge, Oxford, Berlin, Paris, Washington, Rome, Tokyo, Toronto
NATION	China, America, England, UK, Germany, France, USA, Italy, Japan, Spain
PLACE	factory, concert, museum, library, bar, zoo, park, theater, hospital, church
HUMAN	female, male, man, woman, human, boy, girl, elder, gentleman, guys
FURNITURE	chair, desk, table, bed, cabinet, computer, lamp, mirror, house, room

Table 4: Specific concepts from different semantic groups.