Applied Natural Language Processing
Knowledge Based Token Embeddings For Language Models

Yesun-Erdene Jargalsaikhan, Matr. 247523, y.jargalsaikhan@studenti.unitn.it

February 9, 2026

The code discussed in this work and instructions to run it can be found on GitHub.

Abstract

In this project, we explored the development of knowledge-based concept token embeddings to bridge the gap between structured knowledge and LLM representations. We then evaluated the embeddings to verify their ability to reconstruct lexical relations and to assess the feasibility of integrating them with LLMs. The results are promising, ensuring the potential for expanding this methodology and highlighting the importance of dataset richness, embedding models, and architectural design in bridging concept embeddings with large language models.

1 Introduction

Modern Transformer-based language models rely on token embeddings to encode semantic and contextual information. While recent work has shown that optimizing a small set of context embeddings can improve downstream performance, these representations are typically learned without explicit semantic supervision. In contrast, lexical databases encode rich, human-interpretable semantic information. Despite this, relatively little research has explored incorporating such resources into Transformer LLM training. This work aims to bridge the gap between human-defined knowledge and the representations learned by large language models. By injecting concept embeddings derived from lexical databases into Transformer LLMs, we enable explicit control over the semantic structure encoded by the model. Our results shows that human-engineered semantic resources have the possibility of systematically shaping and guiding language model behaviour across downstream tasks.

2 Related Work

Incorporating structured lexical or sense-level information into language models has been an active area of research to enhance semantic understanding. SenseBERT (Levine et al. 2020) improves BERT embeddings by integrating word sense annotations, showing benefits for word sense disambiguation (WSD) and related semantic tasks. Similarly, K-BERT (**kbert**) and LM-KB (**lmkb**) inject knowledge graph information into Transformer models, demonstrating that explicit knowledge can guide LMs to reason over entities and relations. Other approaches, such as ERNIE (**ernie**) and KnowBERT (**knowbert**), combine entity or concept embeddings with contextual representations, improving performance in tasks like entity linking, relation extraction, and semantic disambiguation. These studies collectively highlight the potential of leveraging structured knowledge to shape contextual embeddings, motivating our work on aligning concept-level embeddings with large language models for downstream tasks.

3 Implementation

Our implementation consists of four main stages: (1) constructing the dataset from lexical knowledge bases, (2) learning concept embeddings using Knowledge Graph Embedding (KGE) models, (3) evaluation of the concept embeddings intrinsically and extrinsically to check the feasibility in injection of the concept embeddings to language model representation, and (4) injecting the learned concept embeddings into a language model to assess their effectiveness in a downstream language task, word sense disambiguation.

3.1 Dataset

The dataset is derived from the Universal Knowledge Core (UKC) (Giunchiglia, Batsuren, and Alhakim Freihat 2023), a lexical database of conceptual knowledge, which was used as the foundation

	Train	Validation	Test
Number of triplets	54144	1522	1534
Number of concepts	37123	2475	2481
Number of examples per relations			
has_hyponym	35873	937	926
has_part	10912	371	383
has_attribute	4909	178	177
has_substance	1699	21	26
has_member	751	15	22

Table 1: Entity–Relation Triplet Dataset for KGE (Filtered)

	Train	Validation	Test
Number of triplets	104700	1989	15341946
Number of concepts	88704	3088	3053
Number of examples per relations			
has_hyponym	86185	1482	1504
has_part	11206	253	207
has_attribute	4910	187	167
has_substance	1658	35	49
has_member	741	32	19

Table 2: Entity–Relation Triplet Dataset for KGE (Full)

for the knowledge base and graph embedding. We chose this dataset as it has several layers which covers language independent concepts, where multiple languages to map onto the same underlying concept, whereas WordNet (Miller 1992) and its variants such as WN18RR (Dettmers et al. 2018) is fundamentally language-specific.

3.1.1 Triplet Dataset Construction

We constructed a complete dataset of concept–relation triplets from the UKC knowledge graph, with customized splits for training, validation, and testing. Compared with commonly used benchmarks such as WN18RR, the UKC dataset contains fewer relation types. The final statistics for the training, validation, and test sets are reported in Table 1. For the dataset split, we filtered out triplets containing concept entities that occur fewer than twice, following the methodology used in the construction of other benchmark datasets such as WordNet (Bordes et al. 2014). However, we retained the full dataset for downstream tasks; its reported in the Table 2.

3.2 Knowledge Graph Embedding

Once we created the dataset, we trained several KGE models. Knowledge graph embedding (KGE) aims to learn low-dimensional vector representations of the components of a knowledge graph—entities and relations—while preserving the relational structure of the original graph, for tasks such as link prediction and knowledge graph completion. In this work, we evaluated and exploited the representations of learned concepts of UKC for integration into language models

3.2.1 KGE Models and Training

We considered several state-of-the-art KGE models for embedding the UKC dataset, including TuckER (Balazevic, Allen, and Hospedales 2019), MurP (Balažević, Allen, and Hospedales 2019), RotE, RotH (Chami et al. 2020) and more. The performance of these models is generally comparable; therefore, we trained all models on the UKC dataset. Training is shown in Figure 1

3.2.2 Model Selection and Training Settings

For downstream tasks, the choice of the KGE model is guided by the structural characteristics of the knowledge graph and performance on the dataset, specifically on ranking accuracy measured by Mean Reciprocal Rank (MRR)[1]. Given that the UKC knowledge graph is hierarchical and that the learned representations could subsequently be aligned with an LLM encoder, we adopt embeddings learned via RotE (Chami et al. 2020), a rotation-based Euclidean model.

The best-performing hyperparameter configuration is as follows: a learning rate of 1×10^{-4}, an embedding dimension of 500, a negative sample size of 300, a batch size of 100, and a maximum of 200 training epochs. Training is terminated early using a patience of 5, stopping after five consecutive decrease in mean rank.

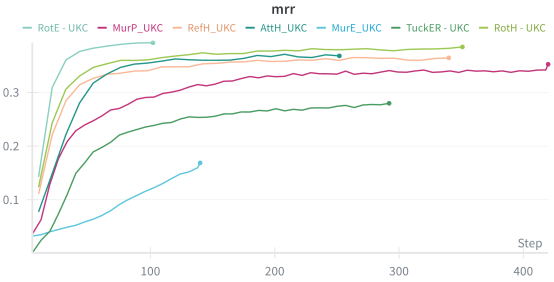


Figure 1: KGE training

[1] Mean Reciprocal Rank (MRR) is a standard evaluation metric for knowledge graph embeddings in link prediction tasks.

Model	Dataset	MRR	hit@10	hit@3	hit@1
RotE	UKC	**0.40**	0.55	**0.44**	**0.31**
RotH	UKC	0.38	**0.56**	0.42	0.29
RefH	UKC	0.36	0.54	0.39	0.27
AttH	UKC	0.37	0.54	0.40	0.28
MurP	UKC	0.35	0.51	0.38	0.27
MurE	UKC	0.16	0.24	0.18	0.12
TuckER	UKC	0.27	0.38	0.30	0.22
RotH	WN18RR	0.49	0.58	0.51	0.44
MurP	WN18RR	0.48	0.56	0.49	0.44

Table 3: Performance metrics from KGE on UKC

4 Evaluation

Intrinsic evaluation measures the quality of the embeddings and their ability to reconstruct and preserve the semantic relations in the knowledge graph, while extrinsic evaluation assesses the alignment between concept embeddings and contextual embeddings produced by the language model.

4.1 Intrinsic Evaluation

We first evaluate the performance of the KGE models on link prediction, and then analysed whether the learned embeddings preserve relational structure by examining geometric relationships inside the embedding space.

4.1.1 KGE metrics

To evaluate the performance of the KGE models and their ability to reconstruct missing head or tail entities from partial triplets, we used standard link prediction metrics, including Mean Reciprocal Rank (MRR)[2] and Hits@K (for $K \in \{1, 3, 10\}$). MRR captures overall ranking quality by averaging the reciprocal of the rank assigned to the correct entity, Hits@K measures the proportion of correct entities ranked within the top K predictions, showing the ability to retrieve related or parent entities. The result of performance metrics of different KGE models on the UKC dataset is summarized in Table 3.

4.1.2 Geometry Inside The Concept Embedding

We evaluated the ability of the embeddings to encode relations between concepts by adopting a formalization proposed in prior work (Park et al. 2025) in which concepts are represented as polytopes, with each vertex corresponding to the vector representation of a concept (e.g., *bird* or *mammal*). Figure 3 shows an example of polytope representations of UKC concepts derived from learned embeddings, which project the unembedding vectors (mapped back into corresponding concept in the unembedding space) on the 3D subspace, with the corresponding hierarchy shown in Figure 2.

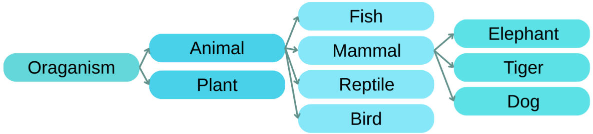


Figure 2: An example hierarchy of concepts in knowledge graph

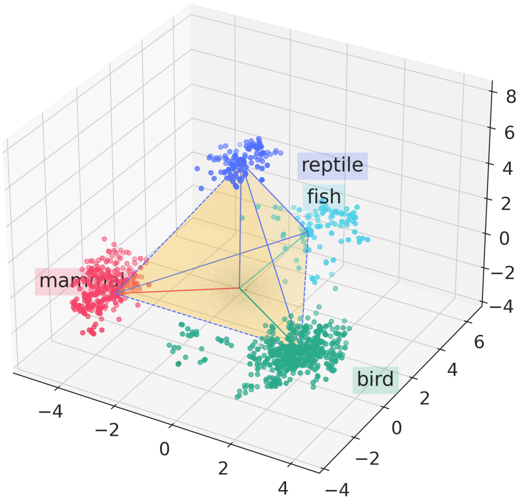


Figure 3: Concepts represented as polytopes.

This formalization allows us to check the semantic hierarchy. We verified that the learned representations of the concepts preserve the hierarchical relations and structure between the concepts. As, it showed in the Figure 4 the semantic hierarchical relation between concepts is encoded geometrically as orthogonality between representations. According to Theorem the work proposed, which implies basically the manipulation of parent concept in the hierarchy should not affect the child concepts.

[2] Mean Reciprocal Rank (MRR) is a standard evaluation metric for knowledge graph embeddings in link prediction tasks.

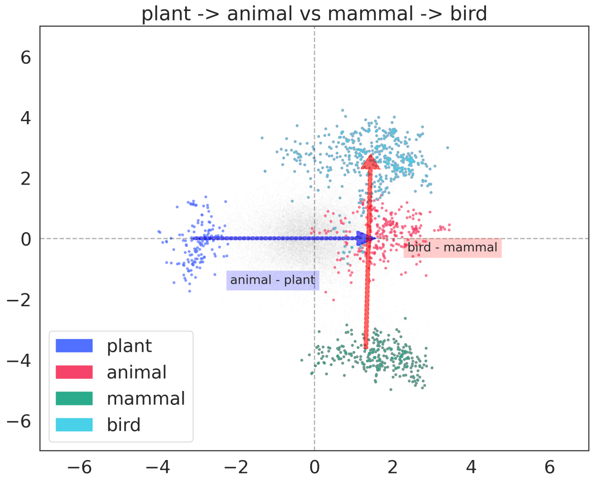
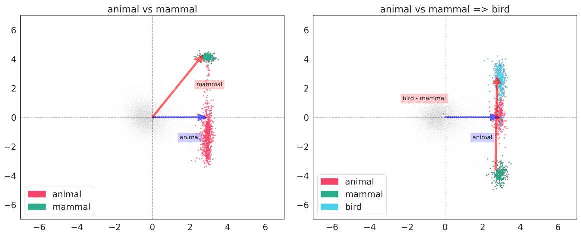


Figure 4: Hierarchical properties inside the concept representations

From the Figure 5a, the vectors showing the span of $\{\bar{\ell}_{\text{animal}}, \bar{\ell}_{\text{mammal}}\}$ which. While in Figure 5b, the span $\{\bar{\ell}_{\text{animal}}, \bar{\ell}_{\text{bird}} - \bar{\ell}_{\text{mammal}}\}$, while the span is directed from the gray points which indicates all the others concepts, and the coloured points belongs to each of the concept



(a) Animal, mammal (b) animal, bird-mammal

Figure 5: Projection of the unembedding vectors of hierarchically related concepts on 2D subspace

Moreover, we showed that hierarchically related concepts are presented as polytopes in the concept representation space, where each vertex is the vector representation of one of the elements in the concepts in Figure 3.

4.2 Extrinsic Evaluation

4.2.1 Mapping from LLM to concept

Here, we evaluate the alignment between learned concept embeddings and contextual representations produced by large language models. We use concept glosses as textual inputs and encode them with multiple pre-trained language models, including BERT (Devlin et al. 2019), mmbert (Marone et al. 2025) and Qwen3 (Zhang et al. 2025). The resulting contextual embeddings are mapped into the concept embedding space, where the ground-truth concept embedding serves as the supervision target. To assess alignment quality, we measure the *geodesic distance*[3] between the predicted concept embedding and the corresponding ground-truth concept embedding. The mapper model has a simple architecture consisting of a linear layer that projects the Qwen embedding dimension (4096) to the RotE embedding space (500), together with a ReLU non-linearity. The architecture of the mapper model is illustrated in Figure 7.

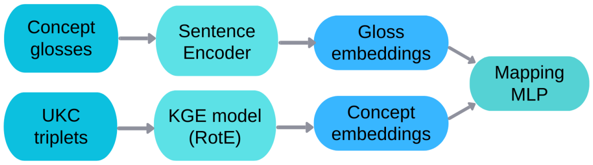


Figure 6: Flow of concept embedding construction and alignment: sentence glosses are mapped via an MLP to the concept embedding space learned by RotE on UKC triplets.

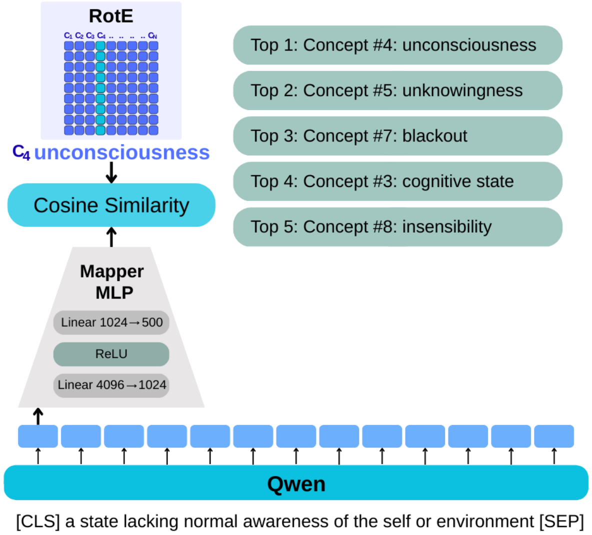

RotE
C₄ unconsciousness
Cosine Similarity
Top 1: Concept #4: unconsciousness
Top 2: Concept #5: unknowingness
Top 3: Concept #7: blackout
Top 4: Concept #3: cognitive state
Top 5: Concept #8: insensibility
Mapper MLP
Linear 1024→500
ReLU
Linear 4096→1024
Qwen
[CLS] a state lacking normal awareness of the self or environment [SEP]

Figure 7: Architecture Mapper Model

As shown in Figure 8, the geodesic distance between the predicted concept and the true concept ranges from 0 to 2, meaning that the predicted concept is either the correct concept or falls within one or two edges away from the ground-truth concept in the knowledge graph. The analyse showed the top K mapped concepts are hierarchically related concepts. These results suggest that the alignment between LLM-encoded concept glosses and the learned concept representations is accu-

[3] A geodesic distance is defined as the shortest path between two points on a graph or manifold.

rate to support their injection into LLMs for downstream tasks.

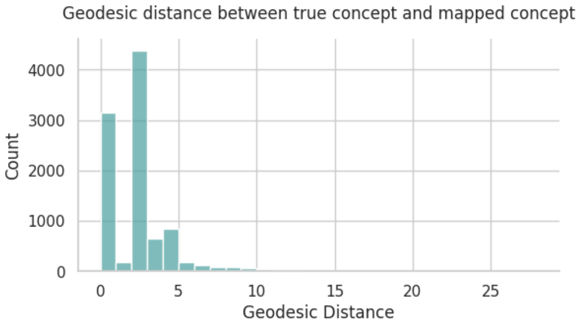


Figure 8: The geodesic distance (count) between true concept and predicted concept

To validate the evaluation, we show the geodesic distance in between random concepts. The result shows that the prediction of exact concept based on it's gloss is high, it predicts the related closer concept or similar concept. As shown in 9, the average distance is around 10 to 12 in random baseline.

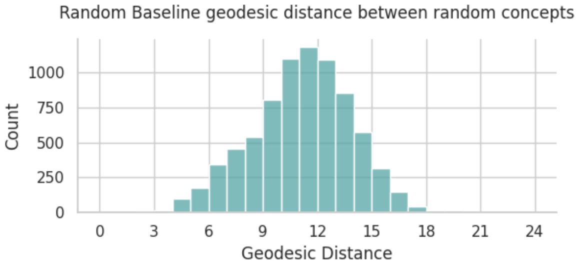


Figure 9: The baseline geodesic distance between random concepts

5 Injecting Concept Embeddings into Language Models

After constructing and evaluating the concept embeddings, we inject them into a language model and evaluate their effectiveness on a downstream task, namely word sense disambiguation (WSD). The goal of WSD is to identify the correct sense—or concept—of a target word given its contextual usage.

5.1 Word Sense Disambiguation

We fine-tune a BERT-based model for the WSD task. Each input sentence is first encoded using a pre-trained multilingual encoder, mmBERT (Marone et al. 2025), to obtain contextualized token representations. From these representations, we extract the embedding corresponding to the target word.

5.1.1 Dataset

For fine-tuning, we train the model on SemCor (Miller et al. 1994), a widely used backbone dataset for word sense disambiguation (WSD). SemCor is a sense-annotated English corpus in which every content word is labeled with a WordNet sense; in our traning, these WordNet senses are mapped to corresponding UKC concept identifiers. For evaluation, we use the SemEval-07 Task dataset (Pradhan et al. 2007), which is commonly used as a benchmark for WSD

5.1.2 Model

To incorporate conceptual knowledge, the target word embedding is linearly projected and combined with the concept embedding matrix via matrix multiplication, producing a score distribution over candidate concepts. The predicted concept is then selected based on this distribution. The architecture of the WSD model is illustrated in Figure 10.

To improve model performance, we apply a concept mask that filters out concepts not observed in the training data from the concept embedding space, thereby reducing noise in the prediction space. To address the strong performance decrease observed on unseen instances, we applied regularization techniques, including label smoothing or temperature scaling on the output logits.

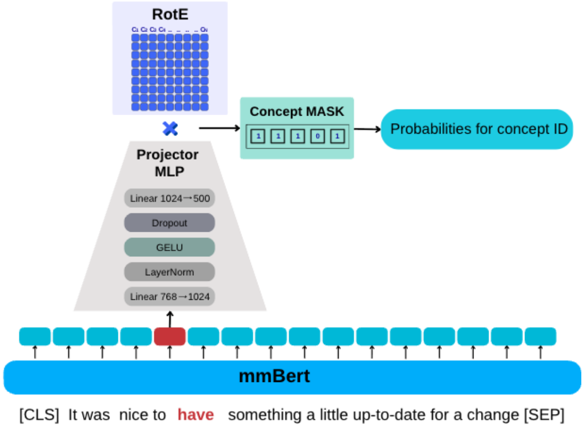


Figure 10: Architecture of the WSD model

5.1.3 Experiment

The best-performing hyperparameter configuration is as follows: learning rate 2×10^{-5}, droupout

Architecture	UKC-WSD	Semantic-WSD
	Precision	
All	74.2	76
Seen	76.4	75.7
Unseen	56.2	77.4
MultiCandidates	66.8	70.4
Multi Seen	72.2	71.3
Multi Unseen	26.3	62.4

Table 4: Result of WSD evaluation

0.1, batch size 16, momentum 0.9, and weight decay 1×10^{-5}. The evaluation results are reported in Table 4. Compared to state-of-the-art knowledge-based WSD approaches, our model underperforms on unseen examples, although it achieves competitive results on seen instances. We expect that performance can be further improved through architectural refinements and more extensive hyperparameter tuning.

6 Results

Our results show that concept embeddings learned from the UKC does preserve lexical and hierarchical relations, with rotation-based KGE models achieving strong link prediction performance. Geometric analysis proves that hierarchical structure is encoded in the embedding space. Extrinsic evaluation shows that contextual embeddings from large language models can be accurately mapped to the concept space, with predicted concepts typically within one or two graph edges of the ground truth. When injected into a language model for word sense disambiguation, the concept embeddings yield has the potential to give better performance on seen instances, showing its effectiveness as an external semantic signal. These results support the feasibility of integrating structured knowledge into LLMs to guide semantic representations.

7 Limitations and Future Work

Our experiments revealed several limitations. The relatively small number of relation types in the UKC dataset, combined with uneven distribution of triplet examples, constrained the performance of the KGE models and prevented higher ranking metrics in link prediction. Expanding the dataset with additional relations and more balanced triplets could improve embedding quality and downstream utility.

Despite limittions and the current model architecture, the learned concept embeddings show potential for integration into large language models, as showed in word sense disambiguation. Future work includes extending this approach to multilingual LLMs, leveraging the UKC's coverage of over 7,000 languages to enable structured semantic knowledge transfer to low-resource languages. Further exploration of mapping architectures, regularization strategies, and downstream tasks could also enhance alignment between concept embeddings and LLM representations.

References

Balazevic, Ivana, Carl Allen, and Timothy Hospedales (2019). "TuckER: Tensor Factorization for Knowledge Graph Completion". 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)*. Association for Computational Linguistics, pp. 5184–5193. DOI: 10.18653/v1/d19-1522. URL: http://dx.doi.org/10.18653/v1/D19-1522.

Balažević, Ivana, Carl Allen, and Timothy Hospedales (2019). *Multi-relational Poincaré Graph Embeddings*. arXiv: 1905.09791 [cs.LG]. URL: https://arxiv.org/abs/1905.09791.

Bordes, Antoine et al. (Feb. 2014). "A semantic matching energy function for learning with multi-relational data". In: *Machine Learning* 94.2, pp. 233–259. ISSN: 1573-0565. DOI: 10.1007/s10994-013-5363-6. URL: https://doi.org/10.1007/s10994-013-5363-6.

Chami, Ines et al. (2020). *Low-Dimensional Hyperbolic Knowledge Graph Embeddings*. arXiv: 2005.00545 [cs.LG]. URL: https://arxiv.org/abs/2005.00545.

Dettmers, Tim et al. (2018). *Convolutional 2D Knowledge Graph Embeddings*. arXiv: 1707.01476 [cs.LG]. URL: https://arxiv.org/abs/1707.01476.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

Giunchiglia, Fausto, Khuyagbaatar Batsuren, and Abed Alhakim Freihat (2023). "One World - Seven Thousand Languages (Best Paper Award, Third Place)". In: ed. by Alexander Gelbukh, pp. 220–235.

Levine, Yoav et al. (2020). *SenseBERT: Driving Some Sense into BERT*. arXiv: 1908.05646 [cs.CL]. URL: https://arxiv.org/abs/1908.05646.

Marone, Marc et al. (2025). *mmBERT: A Modern Multilingual Encoder with Annealed Language Learning*. arXiv: 2509.06888 [cs.CL]. URL: https://arxiv.org/abs/2509.06888.

Miller, George A. (1992). “WordNet: A Lexical Database for English”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. URL: https://aclanthology.org/H92-1116/.

Miller, George A. et al. (1994). “Using a Semantic Concordance for Sense Identification”. In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. URL: https://aclanthology.org/H94-1046/.

Park, Kiho et al. (2025). *The Geometry of Categorical and Hierarchical Concepts in Large Language Models*. arXiv: 2406.01506 [cs.CL]. URL: https://arxiv.org/abs/2406.01506.

Pradhan, Sameer et al. (June 2007). “SemEval-2007 Task-17: English Lexical Sample, SRL and All Words”. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Ed. by Eneko Agirre, Lluís Màrquez, and Richard Wicentowski. Prague, Czech Republic: Association for Computational Linguistics, pp. 87–92. URL: https://aclanthology.org/S07-1016/.

Zhang, Yanzhao et al. (2025). *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*. arXiv: 2506.05176 [cs.CL]. URL: https://arxiv.org/abs/2506.05176.