NEAR^2: A Nested Embedding Approach to Efficient Product Retrieval and Ranking

Shenbin Qian[1,*], Diptesh Kanojia[1], Samarth Agrawal[2], Hadeel Saadany[4], Swapnil Bhosale[1], Constantin Orasan[1] and Zhe Wu[3]

[1]*University of Surrey, United Kingdom*
[2]*eBay Inc, Seattle, WA, USA*
[3]*eBay Inc, San Jose, CA, USA*
[4]*Birmingham City University, United Kingdom*

Abstract

E-commerce information retrieval (IR) systems struggle to simultaneously achieve high accuracy in interpreting complex user queries and maintain efficient processing of vast product catalogs. The dual challenge lies in precisely matching user intent with relevant products while managing the computational demands of real-time search across massive inventories. In this paper, we propose a **N**ested **E**mbedding **A**pproach to product **R**etrieval and **R**anking, called NEAR^2, which can achieve up to 12 times efficiency in model size at inference time while introducing no extra cost in training and improving performance in accuracy for various encoder-based Transformer models. We validate our approach using different loss functions for the retrieval and ranking task, including multiple negative ranking loss and online contrastive loss, on four different test sets with various IR challenges such as *short and implicit queries*. Our approach achieves an improved performance over a smaller embedding dimension, compared to any existing models.

Keywords

ecommerce, search, matryoshka, representation learning

1. Introduction

In e-commerce platforms like Amazon, eBay, and Walmart, effective information retrieval (IR) is crucial for matching user queries with relevant products. However, IR systems face dual challenges of accuracy and efficiency. Accurately interpreting the user intent and ranking search results are complicated by ambiguous, repetitive, and alphanumeric queries [1, 2, 3]. For example, "iPhone 13" often fails to clarify user intent, leading to irrelevant results like "iPhone 13 case" being ranked alongside the intended product. Repetition of query terms in both relevant and irrelevant titles exacerbates this issue. For instance, the term "iPhone 13" might appear in unrelated accessory titles, confusing embedding-based models. Additionally, alphanumeric queries, such as "S2716DG", pose problems because slight variations (*e.g.*, changing "DG" to "DP") signify different product features, which semantic similarity models struggle to interpret without an exact match. These challenges reflect the difficulty of aligning query interpretation with user intent.

At the same time, the computational demands of processing massive product catalogs in real time make efficient retrieval a pressing concern [4]. Balancing accuracy with efficiency remains a significant hurdle for modern IR engines. On the efficiency front, current IR systems often rely on computationally intensive models, such as deep neural networks or large-scale embedding computations, to evaluate semantic similarities between queries and product titles [5, 6]. For instance, calculating embeddings for millions of product titles during a live query can create latency, especially when combined with re-ranking stages that refine results. This latency impacts user experience, as delays of even a fraction of a second

SIGIR-e-Com'25: SIGIR Workshop on eCommerce, 2025, Padua, Italy
*Corresponding author.
s.qian@surrey.ac.uk (S. Qian); d.kanojia@surrey.ac.uk (D. Kanojia); samagrawal@ebay.com (S. Agrawal); hadeel.saadany@bcu.ac.uk (H. Saadany); s.bhosale@surrey.ac.uk (S. Bhosale); c.orasan@surrey.ac.uk (C. Orasan); zwu1@ebay.com (Z. Wu)

can lead to dissatisfaction or abandoned searches. Optimizing these systems to handle large-scale data efficiently without compromising accuracy is a critical challenge in e-commerce search.

In this paper, we propose a **N**ested **E**mbedding **A**pproach to product **R**etrieval and **R**anking, called NEAR^2, which can achieve efficient product retrieval and ranking using much smaller embedding sizes of encoder-based Transformer models [7]. This approach maintains performance comparable to the full model without incurring additional training costs. Our evaluation results on various test sets that contain different types of challenging queries, such as implicit and alphanumeric queries, indicate that NEAR^2 can improve model performance on these challenging datasets using significantly smaller embedding dimension sizes. Our contributions can be summarized as follows:

- We propose NEAR^2, a nested embedding approach, which can achieve up to 12$\times$ efficiency in model size and 100$\times$ smaller in memory usage during inference while introducing no extra cost in training.
- We evaluate NEAR^2 on four different test sets that contains various types challenging queries. Evaluation results show that our approach achieves an improved performance using a much smaller embedding dimension compared to any existing models.
- We conduct ablative experiments on different encoder-based models fine-tuned using different IR loss functions. We find that NEAR^2 is robust to different IR losses or loss combinations for continued fine-tuning.
- We perform a qualitative analysis on retrieved product titles using challenging queries. Our analysis re-affirms the superior performance of our approach and reveals that the similarity scores from NEAR^2 models are more reliable than those of baseline models.

2. Related Work

Modern IR systems encounter several challenges that hinder their performance, particularly in dealing with complex queries and data representation. Ambiguities in natural language, vocabulary mismatches, and the need for scalable real-time processing pose significant challenges [5]. Traditional term-based models often fail due to lexical gaps and polysemy, necessitating the transition to advanced semantic models. Semantic retrieval with dense representations, powered by neural networks and pre-trained language models (PTLMs) like BERT [8], has shown remarkable improvements in handling context and semantics. However, these models demand substantial computational resources and struggle with implicit or alphanumeric queries [5]. Similarly, interaction-based approaches focus on capturing query-document dynamics through deep neural networks, such as the Deep Relevance Matching Model [9], but often sacrifice efficiency and scalability due to their inability to cache document embeddings offline and their reliance on real-time computation [10]. To gap the mismatch of user intent and retrieved product titles in search queries, Saadany et al. [3] curated a dataset annotated with user-intent centrality scores, and proposed a dual loss optimization strategy to fine-tune PTLMs on the dataset in a multi-task learning setting, to solve such challenges.

To address the efficiency issue, researchers have proposed a range of solutions aimed at enhancing efficiency while maintaining accuracy at the same time. Efficiency issues can be tackled through using DUET models that employ local and distributed deep neural networks, which learns dense lower-dimensional vector representations of the query and the document text for efficient retrieval [10]. Knowledge distillation, where smaller models inherit knowledge from larger PTLMs, has proven effective in reducing resource requirements without compromising performance for IR systems [11]. To mitigate computational overhead, Wan et al. [12] proposed to use dimension reduction and distilled encoders to create lightweight models for fast and efficient question-answer retrieval. Kusupati et al. [13] proposed Matryoshka representation learning (MRL) which is able to encode information at different granularities, to adapt to the computational constraints of various downstream tasks. In this paper, we tackle the challenges of accuracy and efficiency using a nested embedding approach based on MRL to create lightweight embedding models for IR tasks.

3. Methodology

This section describes our nested embedding approach in § 3.1 and the backbone models in § 3.2.

3.1. Nested Embedding Training

We utilize MRL with a ranking loss to train nested embeddings of different sizes on various models.

Matryoshka Representation Learning MRL develops representations with diverse capacities within the same higher-dimensional vector by explicitly optimizing sets of lower-dimensional vectors in a nested manner, as illustrated in Figure 1.

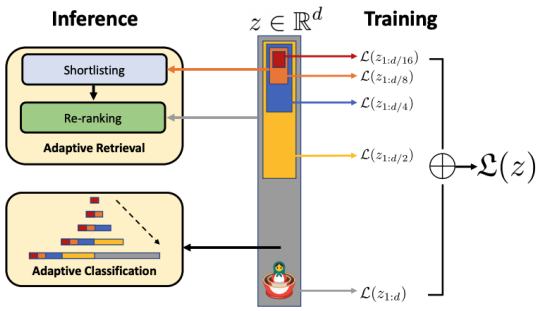


Figure 1: MRL [13] learns multiple nested embedding representations of different sizes ($z \in \mathbb{R}^d$ as the full embedding representation) during training, which are adaptive to different downstream tasks such as retrieval or classification during inference.

The initial $m-$dimensions of the Matryoshka representation, where $m \in M$, the set of nested representation sizes, form a compact and information-dense vector that matches the accuracy of a separately trained $m-$dimensional representation, but requires no extra training effort. As dimensionality increases, the representation progressively incorporates more detailed information, providing a nested coarse-to-fine representation. This approach maintains near-optimal accuracy relative to the full dimensional scale, while avoiding substantial training or deployment costs [14].

The MRL loss is formally defined in Equation 1, where L_{task} is the loss for downstream tasks such as the cross-entropy loss for classification tasks. $f_m(x)$ is the output of the m-th nested embedding representation, and c_m is the importance weight for the m-th embedding representation.

$$L_{MRL} = \sum_{m \in M} c_m L_{task}(f_m(x), y) \tag{1}$$

MRL learns multiple nested embedding representations, each with a different size $m \in M$. The final MRL loss is a weighted sum of the task losses for each of the nested representations. For our product retrieval and ranking task, we set the multiple negative ranking loss (MNRL) [15] as our L_{task}.

Multiple Negative Ranking Loss MNRL measures the difference between relevant (positive) and irrelevant (negative) examples associated with a given query. This technique ensures a clear separation by reducing the distance between the query and positive samples while increasing the distance from negative samples. Using multiple negative examples enhances the model's ability to discern varying levels of irrelevance, refining its optimization. The MNRL objective function is formulated as follows:

$$MNRL = \sum_{i=1}^{P} \sum_{j=1}^{N} max(0, f(q, p_i) - f(q, n_j) + margin) \tag{2}$$

In Equation 2, P represents the number of positive samples; N denotes the number of negative samples; q is the query; f is the similarity metric (cosine similarity in our case), and the $margin$ is a

hyperparameter defining the ideal distance between positive and negative samples based on the relevance score. The goal of MNRL is to minimize the similarity between (q, p_i) while simultaneously maximizing the difference between (q, n_j) for all positive and negative samples.

3.2. Backbone Models

We used encoder-based Transformer models as our backbone for training nested embeddings for efficient product retrieval and ranking.

Pre-trained Language Models We initially leveraged BERT [8], a publicly available pre-trained encoder Transformer model. For our specific use case in e-commerce, we also employed eBERT[1], a proprietary multilingual language model pre-trained internally at eBay. This custom model was pre-trained on a corpus of approximately three billion product titles, supplemented by data from general domain sources like Wikipedia and RefinedWeb.

Expanding our experimental approach, we also incorporated eBERT-siam, a fine-tuned variant of eBERT using a Siamese network architecture. This model aims to generate semantically aligned embeddings for item titles, making it particularly effective for similarity-based search and retrieval tasks. Consistent across all models, we maintained a uniform architectural design of 12 layers with a dimension size of 768.

User-intent Centrality Optimized (UCO) Models Saadany et al. [3, 16] show how current IR systems have problems in achieving user-centric product retrieval and ranking due to implicit or alphanumeric queries. They curated a dataset with user-intent centrality scores (see Section 4.1) and proposed a few models optimized for user-intent using an MNRL loss for retrieval and ranking, and an online contrastive loss (OCL) for user-intent centrality. OCL builds on the traditional contrastive loss (CL) [17] approach but introduces a more focused strategy. While conventional CL uses a twin network to evaluate similarities between all data point pairs from the same and different classes, OCL targets only the most challenging and informative pairs within a batch. By prioritizing such cases, OCL refines the loss calculation to focus on the most critical and complex relationships between data points.

They applied the two losses in a transfer learning setup for eBERT and eBERT-siam models, and performed fine-tuning for centrality classification. Their results indicate that the UCO models achieve an improved performance for retrieval and ranking. Details can be found in Saadany et al. [3].

To improve model efficiency and meanwhile leverage optimized performance of the UCO models, we continued training them using NEAR[2] for both eBERT-UCO and eBERT-siam-UCO models.

4. Experimental Setup

This section explains the datasets we used for training, validating and testing our approach in § 4.1. Implementation details and evaluation metrics are presented in § 4.2 and § 4.3 respectively.

4.1. Data

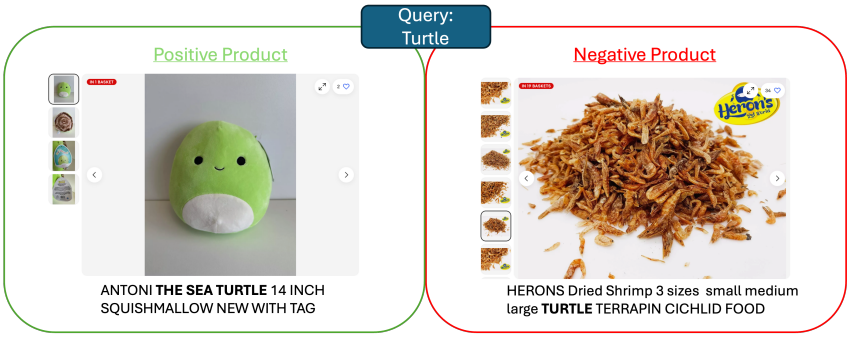
We utilized eBay's internal graded relevance (IGR) datasets to train our nested embedding representation. These datasets comprise user search queries alongside the product titles retrieved on the platform. They are annotated by humans following specific guidelines to generate two types of buyer-focused relevance labels.

The first is a relevance ranking scheme, where query-title pairs are assigned a rank from (1) Bad, (2) Fair, (3) Good, (4) Excellent, to (5) Perfect. A "Perfect" rating signifies an exact match between the query and title, indicating high confidence that the user's needs are fully met, whereas a "Bad" rating indicates no alignment between the query and the product title. This ranking methodology aligns with previous

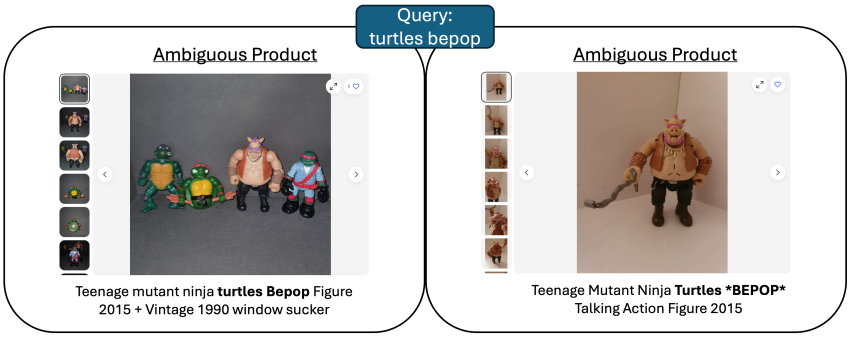
[1] eBERT Language Model

studies [18, 19]. The second annotation type is a binary centrality score, derived through majority voting among multiple annotators, indicating whether a product aligns with the user's expressed query intent. Centrality scoring differs from relevance ranking in that it assesses whether an item is an outlier or unexpected in the retrieval set versus being a core match to user expectations.

To compare the results of our approach with those reported in Saadany et al. [3], we utilized the Common Queries (**CQ**), CQ Balanced (**CQ-balanced**), CQ Common String (**CQ-common-str**), and CQ Alphanumeric (**CQ-alphanum**) test sets proposed in their paper. The CQ test set was constructed using queries with both positive (relevancy > 3) and negative (relevancy < 3) titles, resulting in a dataset skewed toward positive pairs due to the nature of e-commerce data collection. To address this imbalance, a new version, CQ-balanced, was created with approximately equal numbers of positive and negative query-title pairs. The CQ-common-str set was derived by selecting queries where the exact query string appeared in both positive and negative titles, ensuring a strong correlation between relevance scores (both graded relevance and binary centrality). Finally, CQ-alphanum was created to include only query-title pairs containing alphanumeric characters, allowing for a more focused evaluation. Details about their formulation can be found in Saadany et al. [3]. An example of the datasets and the size for each test set can be seen in Figure 2 and Table 1.



(a) The query "turtle" is a part of both positive and negative titles with very different product search outputs. It could also be a part of the ambiguous query "turtles bepop".



(b) The query "turtles bepop" is ambiguous as it could be referred to the major antagonist, "Bepop" or together with other Ninjia Turtles.

Figure 2: Examples of query-title pairs from the *CQ-common-str* test set. The search queries can be very short and ambiguous, but the retrieved products can be very different as shown in (a), or their titles can be quite close in semantic relation as shown in (b).

4.2. Implementation Details

We continued training the PTLMs and the UCO models in § 3.2 for 2 epochs, using our nested embedding approach at dimension sizes of 768, 512, 256, 128 and 64, on the query-title pairs using only the relevance ranking scores (excluding pairs with a score of 3) of the IGR datasets.

During training, we ran a sequential evaluator on the ranking score data to validate for all dimension sizes. First, the evaluator computes the embeddings for both query and title and uses them to calculate

Test Name	# Corpus	# Queries
CQ	187469	17325
CQ-balanced	46561	17325
CQ-common-str	12508	6351
CQ-alphanum	162115	12333

Table 1
The size of the four test sets.

Model	Precision@k			Recall@k			NDCG@k			MRR@k
	3	5	10	3	5	10	3	5	10	10
CQ test										
eBERT-siam	+11.80%	+11.79%	+11.49%	+9.99%	+9.72%	+9.07%	+11.50%	+11.23%	+10.65%	+9.06%
eBERT-UCO	+2.98%	+3.28%	+3.90%	+3.12%	+2.99%	+3.16%	+3.27%	+3.34%	+3.47%	+3.03%
eBERT-siam UCO	+2.82%	+2.75%	+3.16%	+2.72%	+2.45%	+2.50%	+2.91%	+2.77%	+2.80%	+2.58%
CQ-balanced test										
eBERT-siam	+8.85%	+8.45%	+7.31%	+8.85%	+8.43%	+7.28%	+10.28%	+10.03%	+9.56%	+10.48%
eBERT-UCO	+3.19%	+2.87%	+2.42%	+3.15%	+2.81%	+2.41%	+3.36%	+3.19%	+3.03%	+3.25%
eBERT-siam UCO	+2.77%	+2.45%	+2.09%	+2.75%	+2.48%	+2.05%	+3.06%	+2.93%	+2.77%	+3.01%
CQ-common-str test										
eBERT-siam	+6.62%	+4.90%	+3.00%	+6.59%	+4.84%	+3.01%	+8.57%	+7.70%	+6.99%	+8.51%
eBERT-UCO	+1.69%	+1.53%	+0.81%	+1.68%	+1.51%	+0.86%	+1.56%	+1.48%	+1.27%	+1.38%
eBERT-siam UCO	+1.49%	+1.22%	+0.81%	+1.48%	+1.18%	+0.83%	+1.86%	+1.72%	+1.59%	+1.85%
CQ-alphanum test										
eBERT-siam	+5.82%	+5.84%	+6.15%	+4.70%	+4.59%	+5.01%	+5.52%	+5.40%	+5.35%	+4.41%
eBERT-UCO	+3.64%	+3.75%	+3.92%	+3.61%	+3.55%	+3.60%	+3.30%	+3.33%	+3.40%	+2.57%
eBERT-siam UCO	+2.32%	+2.13%	+2.68%	+2.15%	+1.87%	+2.36%	+2.33%	+2.13%	+2.38%	+2.28%

Table 2
Delta in precision, recall, NDCG, and MRR at k on all the test sets for different encoder-based models fine-tuned using **NEAR**^{2} **at** 64 **dimensions** of the entire embedding size (768).

the cosine similarity. Then, it finds the most relevant product title to the query (top 3, 5 and 10 titles) in the corpus of all titles with a max corpus size of $200,000$. For all experiments, we set a batch size of 32, a margin of 0.75 for the MNRL loss with the AdamW optimizer [20] and the learning rate as $5e-05$. Training one model using the above hyperparameters takes about 1.5 hours on a single NVIDIA V100 GPU.

4.3. Evaluation Metrics

We evaluated the model effectiveness through multiple established evaluation metrics including precision, recall, normalized discounted cumulative gain (NDCG) [21] and mean reciprocal rank (MRR).

Precision@k quantifies the ratio of pertinent items within the top-k recommended products, focusing on their individual relevance. Conversely, recall@k assesses the proportion of successfully retrieved relevant items compared to the total number of applicable products, regardless of their positioning. NDCG provides a comprehensive assessment of recommendation quality by analyzing both the relevance and positioning of suggested items. This metric compares the actual recommendation order against an idealized ranking, offering a nuanced evaluation of recommendation performance. MRR focuses on measuring the average ranking position of the first relevant item across different queries. A superior MRR indicates the model's capability to prominently feature highly relevant products, thereby enhancing user experience and recommendation effectiveness.

5. Results and Discussion

We display results achieved using NEAR^2 with a dimension size of 64 in Table 2. Since BERT and eBERT were not fine-tuned on e-commerce data[2], improvement achieved using our approach is huge, as listed in Table A.1 in Appendix A. The values are shown as the percentage of increase (delta) of the evaluation metrics in comparison of those without using NEAR^2 presented in Saadany et al. [3].

[2]eBERT was only pre-trained on e-commerce data.

Comparing results upon using NEAR^2 *vs* existing models, we find that our approach remarkably improves performance on all test sets for all models in § 3.2, even using embeddings with a dimension size of 64, which is $12\times$ smaller in size and $100\times$ smaller in memory usage than the full model.

Model	Dimension	Precision@5	Recall@5	NDCG@5	MRR@10
eBERT-siam	768	+13.33%	+11.77%	+13.10%	+10.20%
	512	+13.35%	+11.87%	+13.16%	+10.30%
	256	+13.26%	+11.68%	+13.05%	+10.19%
	128	+13.10%	+11.37%	+12.80%	+10.16%
	64	+11.79%	+9.72%	+11.23%	+9.06%
eBERT-UCO	768	+4.25%	+4.04%	+4.34%	+3.50%
	512	+4.27%	+3.97%	+4.37%	+3.57%
	256	+4.18%	+3.83%	+4.23%	+3.49%
	128	+3.86%	+3.52%	+3.97%	+3.42%
	64	+3.28%	+2.99%	+3.34%	+3.03%
eBERT-siam-UCO	768	+3.85%	+3.75%	+3.82%	+3.05%
	512	+3.85%	+3.72%	+3.81%	+3.00%
	256	+3.62%	+3.47%	+3.61%	+2.96%
	128	+3.46%	+3.27%	+3.46%	+2.96%
	64	+2.75%	+2.45%	+2.77%	+2.58%

Table 3
Delta in precision, recall, NDCG, and MRR at k on **CQ test** set for different encoder-based models fine-tuned using **NEAR**^{2} for all dimension sizes.

When comparing results of different dimension sizes from the largest (768) to the smallest (64), as shown in Table 3[3] for the **CQ test** set, we discover that the drop in performance is not significant. Embeddings of some smaller dimensions are even slightly better than larger ones. For example, the performance of the eBERT-siam model using NEAR^2 at dimension 512 is slightly better than 768 for precision, NDCG and MRR. This is also true for other models such as BERT, eBERT and eBERT-UCO, which further indicates the effectiveness of our approach for product retrieval and ranking.

To further validate our approach, we qualitatively compared some product titles retrieved with and without NEAR^2. The comparison consistently confirmed the superior performance of our method. Full details are presented in Appendix **??**.

6. Ablation Study

To verify whether continual training using NEAR^2 can help improve performance and efficiency when models are initially trained with other losses, we conducted several experiments using eBERT and eBERT-siam for ablation studies. First, we continued training the models using NEAR^2, which have been fine-tuned using the MNRL and OCL losses respectively to test if our approach works on each of the two individual losses. Second, we tested training these models using the MRL loss first, and then continued fine-tuning on the MNRL and OCL losses in a multi-task learning setting. The results are contrasted with training without using NEAR^2, which are presented as the percentage of increase (delta) in the evaluation metrics in Table 4.

Our ablative results suggest that applying the nested embedding approach to training embeddings with lower dimensions can improve performance for all models fine-tuned using the MNRL or OCL losses for retrieval and ranking, with much obvious improvement on the models trained using the OCL loss. However, models trained with the MRL loss first, then fine-tuned using the MNRL and OCL losses, show slight performance degradation in terms of NDCG and MRR. This suggests that our approach is most effective when used after training the model with an IR task loss first.

[3]BERT and eBERT results are in Table A.2 in Appendix A.

Method	eBERT		eBERT-siam	
	NDCG@5	MRR@10	NDCG@5	MRR@10
MNRL	+4.26%	+3.48%	+2.98%	+2.51%
OCL	+32.09%	+22.50%	+25.86%	+15.66%
MNRL + OCL	+3.34%	+3.03%	+2.77%	+2.58%
MRL: MNRL + OCL	-3.29%	-1.51%	-3.26%	-1.58%

Table 4
Delta in NDCG@5 and MRR@10 on the **CQ test** set for eBERT and eBERT-siam trained using NEAR^2 on different loss functions. We continued training these models using **NEAR^2 at 64 dimensions** of the entire embedding size (768) after they were fine-tuned on the MNRL and OCL losses separately or together (MNRL + OCL). We also trained them on the MRL loss first and then on the MNRL and OCL losses (MRL: MNRL + OCL).

7. Conclusion and Future Work

E-commerce IR systems face the challenge of balancing accurate interpretation of complex user queries with efficient processing of large product catalogs. To address this, we introduced NEAR^2, a nested embedding approach for efficient product retrieval and ranking. NEAR^2 improves accuracy and achieves up to $12\times$ efficiency in model size and $100\times$ smaller in memory usage during inference, without any increase in pre-training costs. Tested across diverse datasets, including short and implicit queries and alphanumeric queries, our method outperforms existing models with smaller embedding dimensions, demonstrating its robustness across challenging evaluation sets, and with efficiency. Our qualitative analysis reinforces the superior performance of our approach, demonstrating that embeddings generated by NEAR^2 models are significantly more reliable than those of baseline models when evaluated based on similarity scores. For future work, we plan to: 1) evaluate our model performance through A/B testing in deployment, 2) leverage internal data to refine larger decoder-based generalist embedding models like NV-embed-v2 [22], and 3) optimize these models using our NEAR^2 approach.

References

[1] S. Li, F. Lv, T. Jin, G. Lin, K. Yang, X. Zeng, X.-M. Wu, Q. Ma, Embedding-based product retrieval in taobao search, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 3181–3189. URL: https://doi.org/10.1145/3447548.3467101. doi:10.1145/3447548.3467101.

[2] K. Keyvan, J. X. Huang, How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3534965. doi:10.1145/3534965.

[3] H. Saadany, S. Bhosale, S. Agrawal, D. Kanojia, C. Orasan, Z. Wu, Centrality-aware product retrieval and ranking, in: F. Dernoncourt, D. Preoţiuc-Pietro, A. Shimorina (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 215–224. URL: https://aclanthology.org/2024.emnlp-industry.17.

[4] D. N. Mhawi, H. W. Oleiwi, N. H. Saeed, H. L. Al-Taie, An efficient information retrieval system using evolutionary algorithms, Network 2 (2022) 583–605. URL: https://www.mdpi.com/2673-8732/2/4/34. doi:10.3390/network2040034.

[5] K. A. Hambarde, H. Proença, Information retrieval: Recent advances and beyond, IEEE Access 11 (2023) 76581–76604. doi:10.1109/ACCESS.2023.3295776.

[6] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint (2023). arXiv:2308.07107.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin,

Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[9] J. Guo, Y. Fan, Q. Ai, W. B. Croft, A deep relevance matching model for ad-hoc retrieval, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 55–64. URL: https://doi.org/10.1145/2983323.2983769. doi:10.1145/2983323.2983769.

[10] B. Mitra, F. Diaz, N. Craswell, Learning to match using local and distributed representations of text for web search, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 1291–1299. URL: https://doi.org/10.1145/3038912.3052579. doi:10.1145/3038912.3052579.

[11] S. Kim, A. S. Rawat, M. Zaheer, S. Jayasumana, V. Sadhanala, W. Jitkrittum, A. K. Menon, R. Fergus, S. Kumar, Embeddistill: A geometric knowledge distillation for information retrieval, 2023. URL: https://openreview.net/forum?id=BT03V9Re9a.

[12] H. Wan, S. S. Patel, J. W. Murdock, S. Potdar, S. Joshi, Fast and light-weight answer text retrieval in dialogue systems, in: A. Loukina, R. Gangadharaiah, B. Min (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 334–343. URL: https://aclanthology.org/2022.naacl-industry.37. doi:10.18653/v1/2022.naacl-industry.37.

[13] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, et al., Matryoshka representation learning, in: Advances in Neural Information Processing Systems, 2022.

[14] X. Li, Z. Li, J. Li, H. Xie, Q. Li, ESE: Espresso sentence embeddings, arXiv preprint (2024). arXiv:2402.14776.

[15] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, R. Kurzweil, Efficient natural language response suggestion for smart reply, arXiv preprint arXiv:1705.00652 (2017).

[16] H. Saadany, S. Bhosale, S. Agrawal, Z. Wu, C. Orăsan, D. Kanojia, Product retrieval and ranking for alphanumeric queries, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 5564–5565. URL: https://doi.org/10.1145/3627673.3679080. doi:10.1145/3627673.3679080.

[17] F. Carlsson, A. C. Gyllensten, E. Gogoulou, E. Y. Hellqvist, M. Sahlgren, Semantic re-tuning with contrastive tension, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=Ov_sMNau-PF.

[18] Y. Jiang, Y. Shang, R. Li, W.-Y. Yang, G. Tang, C. Ma, Y. Xiao, E. Zhao, A unified neural network approach to e-commerce relevance learning, in: Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data, DLP-KDD '19, Association for Computing Machinery, New York, NY, USA, 2019. URL: https://doi.org/10.1145/3326937.3341259. doi:10.1145/3326937.3341259.

[19] D. Kang, W. Jang, Y. Park, Evaluation of e-commerce websites using fuzzy hierarchical topsis based on e-s-qual, Applied Soft Computing 42 (2016) 53–65. URL: https://www.sciencedirect.com/science/article/pii/S1568494616300047. doi:https://doi.org/10.1016/j.asoc.2016.01.017.

[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[21] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: https://doi.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[22] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, W. Ping, Nv-embed: Improved techniques for training llms as generalist embedding models, 2024. URL: https://arxiv.org/abs/2405.17428. arXiv:2405.17428.

A. Additional Figures and Tables

Model	Precision@k			Recall@k			NDCG@k			MRR@k
	3	5	10	3	5	10	3	5	10	10
CQ test										
BERT	+244.88%	+274.90%	+296.75%	+261.89%	+278.42%	+277.64%	+230.75%	+251.93%	+263.34%	+164.96%
eBERT	+185.18%	+198.72%	+196.57%	+204.36%	+202.87%	+185.80%	+180.69%	+191.63%	+190.49%	+124.19%
CQ-balanced test										
BERT	+261.57%	+239.27%	+207.46%	+262.23%	+239.63%	+207.94%	+273.74%	+261.24%	+245.48%	+262.73%
eBERT	+178.60%	+151.73%	+121.33%	+178.84%	+151.59%	+121.06%	+197.78%	+181.60%	+164.85%	+186.77%
CQ-common-str test										
BERT	+230.82%	+206.66%	+171.23%	+230.87%	+206.58%	+171.61%	+238.59%	+226.38%	+210.68%	+226.68%
eBERT	+148.89%	+125.23%	+98.80%	+148.64%	+125.10%	+98.64%	+167.57%	+154.43%	+141.09%	+160.48%
CQ-alphanum test										
BERT	+176.19%	+202.04%	+215.87%	+177.55%	+199.48%	+201.58%	+164.68%	+181.94%	+188.90%	+117.16%
eBERT	+160.04%	+176.97%	+181.05%	+161.56%	+170.52%	+165.06%	+152.21%	+163.12%	+164.35%	+104.15%

Table A.1
Delta in precision, recall, NDCG, and MRR at k on all the test sets for BERT and eBERT fine-tuned using **NEAR**^{2} **at** 64 **dimensions** of the entire embedding size (768).

Model	Dimension	Precision@5	Recall@5	NDCG@5	MRR@10
BERT	768	+286.80%	+296.32%	+265.40%	+170.99%
	512	+287.11%	+296.11%	+265.57%	+171.13%
	256	+286.80%	+295.49%	+264.91%	+170.52%
	128	+284.27%	+291.95%	+262.16%	+169.54%
	64	+274.90%	+278.42%	+251.93%	+164.96%
eBERT	768	+192.17%	+197.60%	+185.11%	+119.88%
	512	+192.41%	+197.45%	+185.24%	+120.00%
	256	+192.17%	+196.98%	+184.72%	+119.50%
	128	+190.26%	+194.32%	+182.58%	+118.71%
	64	+183.19%	+184.16%	+174.59%	+114.99%

Table A.2
Delta in precision, recall, NDCG, and MRR at k on **CQ test** set for BERT and eBERT models fine-tuned using **NEAR**^{2} for all dimension sizes.

Method	Retrieved Title	Ranking	Sim_Score$_{Norm}$
NEAR^2@64	CRAZY DAISY Shasta daisies Qty 2 PLANTS Hardy Perennial Healthy plants	1	0.3967
	CRAZY DAISY Shasta daisies Qty 2 x Hardy Perennialhealthy plants	2	0.3824
	Streptocarpus MKsArktur09 young plant	3	0.3822
	Spathiphyllum Peace Lily Indoor Plants 1 x Potted Lily House Plant 9cm Pot	4	0.3731
	Houseplant and Pot Package	5	0.3723
	Spathiphyllum Peace Lily House Plant Live Indoor House Potted Tree In 9cm	6	0.3710
	Boston FernLive 10 Plants Lots Of Roots Air Purifier Reptile Terrarium ORGANIC	7	0.3696
	1 x CRAZY DAISY Shasta daisies Hardy Perennial Healthy plant	8	0.3671
	Leucanthemum Crazy Daisy Middleton Nurseries Flowering hardy Plants	9	0.3642
	Syngonium White Butterfly Arrowhead Goose Foot Plant House Plant Easy Care	10	0.3640
W/o NEAR^2@64	Houseplant and Pot Package	1	0.2665
	Spathiphyllum Peace Lily Indoor Plants 1 x Potted Lily House Plant 9cm Pot	2	0.2425
	Spathiphyllum Peace Lily House Plant Live Indoor House Potted Tree In 9cm	3	0.2417
	Cordyline Kiwi Ti Plant 7c Best Indoor Plants 7c Colourful 3040cm Potted Plant	4	0.2349
	68 Live Snake Plant Sansevieria Trifasciata Two Plants	5	0.2341
	Leucanthemum Crazy Daisy in plant in 13cm pot approx	6	0.2338
	Multi Listing Pond Plants Marginal Plants Water Bog Garden Oxygenator SALE	7	0.2317
	12 Succulent Flowers not Included Pots 12 Pcs 12 Fashion Practical	8	0.2267
	Avocado plant	9	0.2255
	3CM Succulent Cactus Live Plant Copiapoa Tenuissima Chile Home Garden Rare Plant	10	0.2239
Gold label	Aloe Vera Plant - Large Plant in Pot	/	/

Table A.3
Retrieved titles for the detailed query “plants” using or not using NEAR^2@64 on eBERT-siam.

Method	Retrieved Title	Ranking	Sim_Score$_{Norm}$
NEAR^2@64	Moonstone Opal Pendant 925 Sterling Silver Necklace Chain Womens Jewellery Gifts	1	0.3934
	Green Triplet Fire Opal Peridot 925 Sterling Silver Jewelry Pendants 27 v957	2	0.3814
	Moonstone Opal Pendant 925 Sterling Silver Necklace Earring Women Jewellery Gift	3	0.3664
	Sterling Silver 925 Signed Opal Heart Pendant Necklace 19 Chain	4	0.3500
	Vintage Possibly Opal Pendant On Gold Tone Necklace Chain	5	0.3337
	Triplet Fire Opal Peridot Gemstone 925 Silver Jewelry Necklace 18 AQ269	6	0.3309
	BULK LOT Vintage 925 Silver Costume Jewellery Gemstones Opal Cloisonne Etc	7	0.3227
	Ethiopian Opal 925 Sterling Silver Choker Necklace Women Gemstone Jewelry Gift	8	0.2829
	Yellow Triplet Fire Opal Citrine 925 Sterling Silver Jewelry Earrings 21 s558	9	0.2691
	Blue Opal Pendant 925 Sterling Silver Minimalist Necklace Gift for Girlfriend	10	0.2688
W/o NEAR^2@64	Vintage Possibly Opal Pendant On Gold Tone Necklace Chain	1	0.2615
	Green Triplet Fire Opal Peridot 925 Sterling Silver Jewelry Pendants 27 v957	2	0.2561
	Moonstone Opal Pendant 925 Sterling Silver Necklace Chain Womens Jewellery Gifts	3	0.2558
	Triplet Fire Opal Peridot Gemstone 925 Silver Jewelry Necklace 18 AQ269	4	0.2505
	Moonstone Opal Pendant 925 Sterling Silver Necklace Earring Women Jewellery Gift	5	0.2475
	Sterling Silver 925 Signed Opal Heart Pendant Necklace 19 Chain	6	0.2472
	GemporiaGems TV Sterling Silver 157ct Ethiopian Blue Opal Pendant Necklace	7	0.2448
	NWT GEMPORIA GEMS TV AUSTRALIAN OPAL STERLING SILVER PENDANT	8	0.2381
	Vintage 925 Silver Opal Ring size J	9	0.2360
	Australian Triplet Opal Gemstone 925 Sterling Silver Handmade Ring All Size	10	0.2270
Gold label	925 Sterling Silver Red Coral Gemstone Handmade Jewelry Vintage Pendant S120	/	/

Table A.4
Retrieved titles for the detailed query “925 sterling silver triplet opal gemstone jewelry vintage pendant s-1.20” using or not using NEAR^2@64 on eBERT-siam.

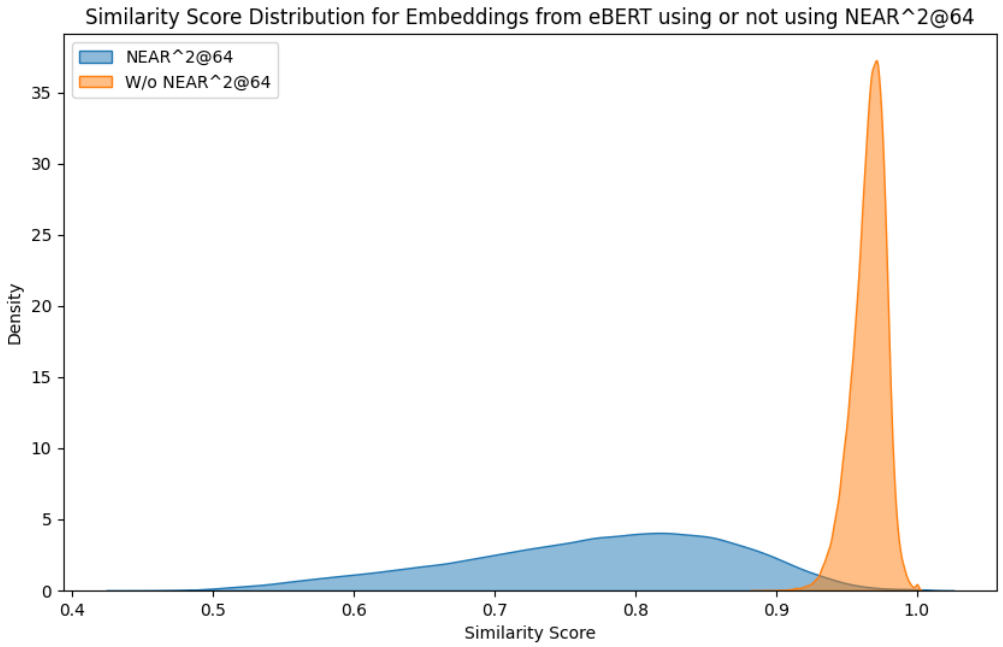


Figure A.1: Similarity score distribution for embeddings from models **using** *vs* **not using** NEAR^2@64 with eBERT on the CQ test set.