Embeddings Might Be all You Need: Domain-Specific Sentence Encoders for Latin American E-Commerce Questions

Rodrigo Caus[1,2][a], Victor Sotelo[1,2][b], Victor Hochgreb[2][c] and Julio Cesar dos Reis[1][d]
[1]*Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil*
[2]*GoBots Company, Campinas, São Paulo, Brazil*

Keywords: Sentence Embeddings, Question Retrieval, E-Commerce.

Abstract: In Latin American e-commerce, customer inquiries often exhibit unique linguistic patterns that require specialized handling for accurate responses. Traditional sentence encoders may struggle with these regional nuances, leading to less effective answers. This study investigates the application of fine-tuned transformer models to generate domain-specific sentence embeddings, focusing on Portuguese and Spanish retrieval tasks. Our findings demonstrate that these specialized embeddings significantly outperform general-purpose pre-trained models and traditional techniques, such as BM-25, thereby eliminating the need for additional re-ranking steps in retrieval processes. Our results investigate the impact of multi-objective training within Matryoshka Representation Learning, demonstrating its effectiveness in maintaining retrieval performance across various embedding dimensions. Our approach offers a scalable and efficient solution for multilingual retrieval in e-commerce, reducing computational costs while ensuring high accuracy.

1 INTRODUCTION

In the rapidly growing e-commerce landscape, effective customer service through accurate question-answering systems is crucial to user satisfaction and conversions. Sentence encoders (Reimers and Gurevych, 2019) play a central role in these systems, capturing semantic meaning, context, and relationships in numerical embeddings. Such embeddings can be used to select the most appropriate answer to the customer inquiry.

General-purpose sentence encoders often prove less effective in specialized domains due to their difficulty in capturing unique vocabulary, phrasing, and contextual nuances (Tang and Yang, 2025). This entails that generic models frequently require high-dimensional embeddings and separate re-ranking models to achieve acceptable domain-specific effectiveness, especially when resource minimization is a key objective.

GoBots[1] company addresses a high volume of customer inquiries from e-commerce platforms in Spanish and Portuguese. We have implemented an end-to-end question-answering solution based on embeddings to manage customer queries. Existing pretrained solutions assist in retrieving suitable text (questions) to provide answers. This context requires performing a re-ranking process to ensure the quality of the retrieved text (Chico et al., 2023).

However, this multi-component approach inherently increases complexity and can compromise the overall quality and efficiency of the retrieval pipeline. Employing a distinct retriever and a subsequent re-ranker directly escalates computational resource demands, which is prohibitive for small business scenarios. Such an architecture typically requires significantly more memory and CPU processing per query, leading to higher operational costs and potentially impacting end-user response latency. In contrast, fine-tuning domain-specific sentence encoders may offer a more direct path to optimize cost, processing, and storage.

This study investigates resource optimization strategies for e-commerce question paraphrase re-

[a] https://orcid.org/0000-0002-0904-4865
[b] https://orcid.org/0000-0001-9245-8753
[c] https://orcid.org/0000-0002-0529-7312
[d] https://orcid.org/0000-0002-9545-2098

[1]Leading company of artificial intelligence (AI) solutions for the e-commerce sector in Latin America. Official website: https://gobots.ai.

Caus, R., Sotelo, V., Hochgreb, V. and Reis, J. C.
Embeddings Might Be all You Need: Domain-Specific Sentence Encoders for Latin American E-Commerce Questions.
DOI: 10.5220/0013821200004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 135-146

trieval pipelines that integrate vector-based retrieval (potentially utilizing dense or sparse vectors) with a subsequent re-ranking phase. This research aims to attain two specific goals:

1. Assess the feasibility and effectiveness of utilizing a single, unified embedding model to generate representations for retrieval pipelines, comparing their performance against conventional two-model architectures (*i.e.*, separate models for retrieval and re-ranking).
2. Analyze the trade-off between the reduction in embedding dimensionality from such a unified model and the consequent impact on retrieval effectiveness and computational efficiency.

Our findings demonstrate that a single, domain-fine-tuned embedding model, trained efficiently on a single, commonly available GPU, outperforms the multi-model encoder-re-ranker pipeline and BM-25 retrieval in a real-world e-commerce setting. This study, conducted in collaboration with a company, highlights the practical benefits of this streamlined approach.

As key contributions, we release our test and calibration datasets. Notably, these datasets are in Portuguese and Spanish, which are often underrepresented in natural language processing research, offering valuable resources for extending existing embedding model benchmarks, such as MTEB (Enevoldsen et al., 2025). Furthermore, we are open-sourcing our training and validation code, enabling other researchers and practitioners to adapt and apply these methods to their domains[2].

The remainder of this article is organized as follows: Section 2 presents a synthesis and analysis of key related studies. Section 3 summarizes the E-FAQ, a dataset generated in our research. Section 4 describes the training details and approaches used in this research. Section 5 outlines our experimental evaluation, which includes the dataset, baselines, and evaluation metrics. Section 6 reports on our results obtained. Section 7 discusses our findings. Finally, Section 8 summarizes the conclusions and suggests directions for future research.

2 RELATED WORK

Related work, in the context of our research, concerns training domain-specific and language-specific embedding models, particularly for information retrieval tasks.

[2]Available at https://github.com/rodrigocaus/embedding-training.

On domain-specific embedding models, Z. Feng et al. (Feng et al., 2020) introduced *CodeBERT*, a transformer-based model trained on open-source GitHub repositories, which currently supports only six programming languages. It follows multilingual BERT approaches, using masked language techniques during fine-tuning. The models focus on bimodal data, aligning text (code documentation) with their respective code during pre-training. After this initial training, they utilize the base model to fine-tune the process, thereby improving the alignment between text and code representations. They test the effectiveness of code retrieval based on natural language queries, and *CodeBERT* outperformed results from other pre-trained models, such as RoBERTa, achieving a higher Mean Reciprocal Rank in the CodeSearchNet benchmark.

Clinical BERT (Alsentzer et al., 2019) models were developed to meet the need for domain-specific embeddings in clinical contexts. The authors initialized Clinical BERT using two primary models: Base BERT and BioBERT. They followed the same training procedures used for BERT, utilizing a corpus of clinical texts. Their findings showed that specialized domain models performed better in domain classification tasks for clinical benchmarks. However, a limitation of these models is their limited generalization to datasets that differ from the training data.

Regarding language-specific embedding models, Huang et al. (Huang et al., 2024) introduced Piccolo 2, a state-of-the-art model on Chinese embedding benchmarks. It leverages an efficient multi-task hybrid loss training approach, effectively leveraging textual data and labels for various downstream tasks, combined with Matrioshka Representation Learning (MRL) to support more flexible vector dimensions. It was evaluated over six tasks on CMTEB benchmark, including text retrieval, pair classification, and semantic similarity.

Industrial Applications models (Bednář et al., 2024) focused on creating an embedding with a lower size to improve computational efficiency. They applied the study to Seznam, a Czech search engine, and explored techniques suitable for non-English languages, utilizing datasets from non-public sources. Their study examined three methods: auto-encoder training, unsupervised contrastive fine-tuning, and multilingual distillation, which do not require large datasets, making them practical for real-world use. The models were evaluated on semantic textual similarity (STS) and COSTRA, a benchmark for assessing embedding quality, as well as measuring search engine ranking effectiveness using precision at 10. Their findings showed that pretrained versions and multi-

lingual distillation provide the best encoder models, highlighting their effectiveness in enhancing search result quality.

DeepFAQ (Chico et al., 2023) refers to a Portuguese automatic question-answering system that uses semantic search to find similar questions from a database of FAQs. Its solution applies a general domain embedding to represent the data (questions and answers). It retrieves candidate questions and applies a domain-specific re-ranking model to identify the most relevant one, ultimately providing the corresponding answer.

Our current approach makes a novel and original contribution by utilizing domain-specific embeddings for the e-commerce sector, specifically tailored for Brazilian Portuguese and Spanish, two low-resource languages in NLP. We take advantage of the approach of language-specific embedding presented by Huang et al. (Huang et al., 2024) to fine-tune sentence encoding models. These embeddings effectively capture the nuances of informal language used on online platforms, thereby enhancing results in e-commerce-related NLP tasks and addressing gaps identified in previous methods, particularly the encoder-re-ranker pipeline as presented by Huang et al. (Chico et al., 2023).

3 E-FAQ: GROUPED FREQUENTLY ASKED QUESTIONS FROM E-COMMERCE

Real-world data are fundamental for generating domain-specific sentence embeddings. This section presents the E-FAQ, a weakly-supervised dataset of e-commerce frequently asked questions (FAQs), with sentences uttered in Brazilian Portuguese or Spanish. Each entry i of the dataset is the tuple $(q_i, \mathbf{S}_i, \mathbf{A}_i, \mathbf{D}_i)$, in which:

- q_i is an anchor question sentence.
- $\mathbf{S}_i$ is a set of sentences that are similar to q_i; the sentences convey the same meaning and are interchangeable with q_i.
- $\mathbf{A}_i$ is a set of sentences that are almost similar to q_i; the sentences are closely related to q_i, but differ in meaningful detail.
- $\mathbf{D}_i$ is a set of sentences that are dissimilar to q_i; the sentences discuss different topics or contain unrelated information with q_i.

Any of $\mathbf{S}_i$, $\mathbf{A}_i$ or $\mathbf{D}_i$ sets can be empty for a given i. However, at least one of the sets is not empty. Figure 1 illustrates data examples from E-FAQ. For instance, relative to the anchor question about a backpack's volume, a query about a "bag" is considered 'Similar', while a related but distinct question about fitting a "laptop" is labeled 'Almost Similar'. This highlights the semantic nuances our methodology is designed to capture, distinguishing between identical intent, related topics, and entirely dissimilar queries.

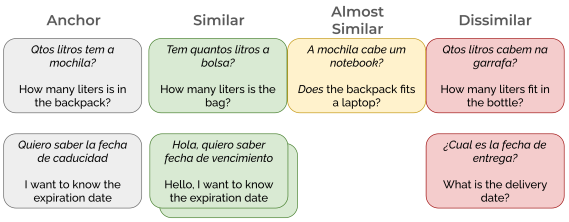

Figure 1: Examples of entries in E-FAQ. The figure illustrates our classification scheme, where candidate questions are labeled as 'Similar', 'Almost Similar', or 'Dissimilar' with respect to an "Anchor" question. Note the subtle distinction between 'Similar' (same intent, e.g., "backpack" vs. "bag") and 'Almost Similar' (related topic, different intent, e.g., backpack capacity vs. fitting a laptop), which allows for a more nuanced understanding of semantic relevance. All sentences are uttered in Brazilian Portuguese or Spanish. English translations are presented below the original sentences.

We originally created this dataset to address a resource-scarce gap for Portuguese and Spanish, particularly within the e-commerce domain. We gathered questions from Latin American e-commerce websites sourced from the *GoBots* database (cf. Figure 2). Initially, we collected a larger set of questions; after removing duplicates and questions containing fewer than four words, we were left with one million questions, evenly split between Brazilian Portuguese and Spanish.

Our primary goal with E-FAQ was to construct a dataset composed of thematically disjoint question groups. To achieve this, we followed a structured three-step pipeline: (1) a Natural Language Understanding (NLU) analysis for feature extraction; (2) a clustering phase to group related questions and filter noise; and (3) an intra-cluster classification stage to assign fine-grained similarity labels. Each of these steps is detailed in the subsequent sections.

3.1 Natural Language Understanding

In the first stage of our pipeline, we employed natural language understanding (NLU) models to extract intents and named entities from each question. For this, we leveraged a proprietary machine learning model previously trained on a large corpus of sentences within the *GoBots* data environment. The model performed two key tasks. First, it identified the

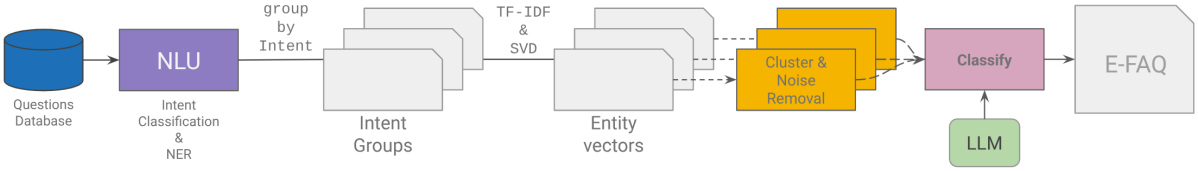

Figure 2: Overview of data collection process in generating the E-FAQ dataset.

user's intent — the overall purpose of the question — which allowed us to classify each query into one of 64 distinct thematic categories. Second, it recognized named entities — a term or expression with a known meaning relevant to the sentence's comprehension — which were used to normalize the text by mapping synonymous terms to a canonical form and correcting potential typos.

3.2 Clustering

Within each intent category, we employed the HDBSCAN clustering algorithm to group similar questions. We used the extracted entities as the main features, as they correspond to a normalized sequence of relevant terms of the sentences. We used the 95% most frequent entities to create a TF-IDF sparse representation, and then applied a singular value decomposition (SVD) to reduce the features dimension.

We specifically chose HDBSCAN for its ability to group semantically similar questions into dense clusters while simultaneously identifying and filtering out noisy data—i.e., questions that do not belong to any coherent group, ensuring that the resulting clusters are thematically distinct from one another. We consider clusters with at least two sentences. This process yielded more than 142,000 clusters, encompassing over 445,000 examples, with the cluster medoid serving as the anchor sentence.

3.3 Classification

In the final step, we analyze the contents of each cluster to label every question relative to its cluster's anchor sentence, to ensure high-quality semantic similarity data. This in-cluster classification step was conducted with a synthetic labeling process, using large language models as annotators.

To secure the high value of the annotation process, we curated a calibration dataset, which began with an initial pool of 150 real-world question pairs sourced from e-commerce platforms. Each pair was independently classified as "similar", "almost similar", or "dissimilar" by three human annotators: two computer science graduate students with expertise in AI for e-commerce (both co-authors) and one undergraduate student without prior experience in the domain. Each annotator was presented to the instruction:

> *You'll see pairs of product questions extracted from e-commerce platforms. Your task is to label these pairs according to their semantic similarity. The label will be one of:*
>
> - ***similar**: The sentences convey the same meaning or idea, even if phrased differently. For e-commerce, these questions could be answered with the same answer.*
> - ***almost similar**: The sentences share a significant amount of information and are strongly related, but there are subtle differences in meaning or scope. They are similar, but cannot be answered with the same answer.*
> - ***dissimilar**: The sentences contain distinct information or completely different meanings, and are not correlated.*

We established the final label for a pair based on majority vote. Of the initial 150 pairs, 144 reached a majority consensus (96%), meaning at least two annotators agreed on a label. The remaining 6 pairs, for which each annotator assigned a different label, were therefore discarded. The resulting calibration dataset, which we named *GoSim3*, is available on the Hugging Face Hub[3].

We then leveraged this calibration dataset to optimize a classification LLM prompt. Specifically, we used the Gemma 3 language model (Kamath et al., 2025) to identify the prompt that yielded the highest accuracy against the *GoSim3* ground-truth labels. The prompt contained the same instructions given to the human annotators, as other formatting and reasoning instructions. This optimized prompt and model were subsequently used to classify each question pair within our 142,000 clusters, ensuring a reliable, large-scale assessment of semantic similarity.

3.4 Dataset Split

The dataset was further divided into *training*, *validation*, and *test* sets. The training set comprised most of

[3]Available at https://huggingface.co/datasets/GoBotsAI/GoSim-3.

the data, with 121,248 entries, followed by the validation set, with 13,472 entries. The test sets were organized by language (Portuguese and Spanish) and stratified by intent class, resulting in two sets with 4,000 entries each. For commercial reasons, the training dataset used in this study cannot be publicly released. Nevertheless, our test dataset is available at HuggingFace's Hub[4].

4 TRAINING METHODS

The primary application of our proposed models is retrieving similar questions given an input query. Recent research has increasingly focused on bi-encoder architectures for generating sentence embeddings. These models independently encode the query and the questions, allowing for efficient similarity scoring (Izacard et al., 2022). Formally, given two sentences x and y, their embeddings are generated independently by the f_θ and f_γ models, respectively. The embedding space similarity of the two sentences ϕ can be defined as:

$$\phi(x,y) = cos(f_\theta(x), f_\gamma(y))/\tau \quad (1)$$

In which τ is a temperature parameter. Two transformer models can be used to embed sentences in f_θ and f_γ, as in DPR (Karpukhin et al., 2020), which employs two BERT encoders to map questions and passages into a shared semantic space. Recent studies used a single transformer model f_θ in a siamese bi-encoder architecture to embed the sentences. Figure 3 presents this architecture. Models that use this architecture, like SBERT (Reimers and Gurevych, 2019), LaBSE (Feng et al., 2022), and E5 (Wang et al., 2024a; Wang et al., 2024b), proved to be effective in many zero-shot natural language tasks. As questions and queries share the same domain, we employ the Siamese architecture. For the pooling strategy, we use the mean of the token representations.

We assume that E-FAQ contains disjoint groups of similar sentences, so each dataset entry contains a unique group of questions. Leveraging the "similar", "almost similar", and "dissimilar" labels. We designed a training regimen incorporating two distinct objectives: a retrieval objective and a semantic similarity objective. This multi-task learning strategy allowed the model to simultaneously learn effective representations for retrieving relevant questions and accurately assessing the degree of semantic relatedness between question pairs within our refined

[4]Available at https://huggingface.co/datasets/GoBotsAI/e-faq.

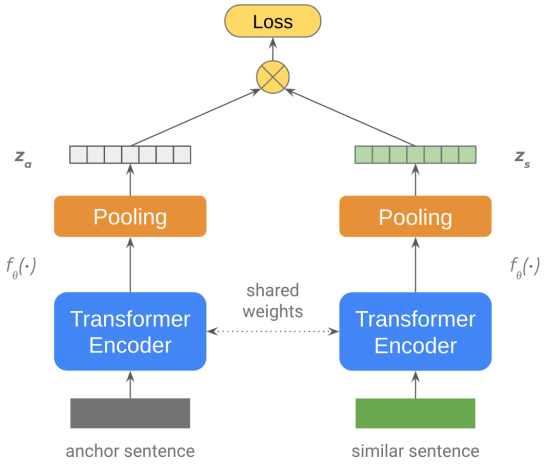

Figure 3: Siamese Dual Encoder model for sentence embeddings generation.

dataset. This method follows Huang et al. (Huang et al., 2024) approach.

For the retrieval objective, we used the InfoNCE loss (van den Oord et al., 2019), in which an anchor question q_i, associated with a similar question s_i, is compared against $N-1$ dissimilar questions in a cross-entropy function. The loss is defined by:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{\phi(q_i,s_i)}}{e^{\phi(q_i,s_i)} + \sum_{j=1,j\neq i}^{N} e^{\phi(q_i,s_j)}} \quad (2)$$

This loss encourages similar question pairs to have higher similarity scores, and dissimilar questions to have lower scores (Izacard et al., 2022).

We define $s_{ij} \in \mathbf{S}_i$ as a question extracted from the set of questions similar to q_i. The training data consisted of entries in the form (q_i, s_{ij}), with $0 \leq j \leq |\mathbf{S}_i|$, augmented from each cluster from E-FAQ. Additionally, we incorporated challenging negative examples by selecting K "hard-negatives" through a combination process from the union of $\mathbf{A}_i$ and $\mathbf{D}_i$. These K hard negatives were then combined with the in-batch negative samples, such that the total number of negative examples considered for each positive sample was $N-1$, where N is the batch size. For entries that yielded fewer than K hard negatives, this set was supplemented by sampling from the hard negatives of other topically disjoint entries from the entire dataset.

The final contrastive loss is a combination of both the original loss function $\mathcal{L}_{ce}$, considering the cross-entropy on anchor sentences, and its symmetric version $\mathcal{L}'_{ce}$, considering the cross-entropy on similar sentences:

$$\mathcal{L}_r = \mathcal{L}_{ce} + \mathcal{L}'_{ce} \quad (3)$$

For the semantic similarity objective, we converted the "similar", "almost similar", and "dissimi-

lar" labels into score values. The training data consisted in triples in the form (q_i, p_{ij}, z_{ij}), in which q_i is the anchor question, p_{ij} is a sentence in q_i's cluster, and z_{ij} is their labeled similarity score, with values:

$$z_{ij} = \begin{cases} 1, & \text{if} \quad p_{ij} \in \mathbf{S}_i \\ 0, & \text{if} \quad p_{ij} \in \mathbf{A}_i \\ -1, & \text{if} \quad p_{ij} \in \mathbf{D}_i \end{cases} \tag{4}$$

We used the Cosine Sentence Loss (CoSENT) (Su, 2022) in this task, a ranking loss function specifically designed for the score-labeled text pairs (Huang et al., 2024). The loss is defined by:

$$\mathcal{L}_s = log\left(1 + \sum_{z_{ij} > z_{kl}} e^{\phi(q_k, p_{kl}) - \phi(q_i, p_{ij})}\right) \tag{5}$$

Equation 6) defines the final multi-task loss:

$$\mathcal{L} = \begin{cases} \mathcal{L}_r, & \text{if task is retrieval} \\ \mathcal{L}_s, & \text{if task is semantic similarity} \end{cases} \tag{6}$$

To achieve our objective of reducing embedding dimensionality, we employed Matryoshka Representation Learning (MRL) (Kusupati et al., 2024) during model training. This technique compels the model to produce hierarchical, coarse-to-fine embeddings, ensuring that these lower-dimensional representations are at least as accurate as independently trained low-dimensional representations.

5 EXPERIMENTS

For our experiments, we fine-tuned two multilingual transformer models. The first, XLM-RoBERTa (Conneau et al., 2020), serves as a strong baseline due to its extensive pre-training on multilingual text. The second, Multilingual E5-Base (Wang et al., 2024b), was selected for its state-of-the-art performance in dense retrieval tasks, as evidenced by its high ranking on the MTEB leaderboard[5]. Both models produce embeddings with a native dimensionality of 768.

All models were trained on a single NVIDIA RTX 4090 GPU using the AdamW optimizer. We set the learning rate to 2×10^{-5} with a linear warmup for the first 10% of training steps, followed by a stable learning rate. We trained for a maximum of $5,000$ steps using a batch size of 256 sentence pairs. The temperature parameter τ for the contrastive loss was fixed at 0.05 to facilitate the discrimination of negative samples. For MRL, we trained the models to produce nested embeddings at dimensions of $\{64, 128, 256, 384, 512, 768\}$. We evaluated the model on a held-out validation set every 200 steps and saved the checkpoint with the highest retrieval accuracy.

For our assessments, we evaluated our trained models primarily on a symmetric retrieval task, specifically sentence paraphrase mining, using the test partition of our domain-specific E-FAQ dataset (cf. Section 3). This dataset, comprising $8,000$ e-commerce queries in Portuguese and Spanish, enables us to directly measure the model's effectiveness in identifying semantically equivalent questions, a core function for customer service applications. This primary task serves as the main benchmark for retrieval performance.

To ensure the embeddings offer in-domain generality and clear similarity separability, we conducted a secondary evaluation on a Semantic Textual Similarity (STS) task. For this, we utilized the *GoSim3* dataset, a domain-specific benchmark that was intentionally excluded from our model's training distribution. This test validates the correlation between human annotations and results obtained by computing the similarity between the vector representations of both questions. It assesses that the model can robustly generalize to new, unseen data within the e-commerce domain and accurately distinguish between varying degrees of semantic relatedness.

5.1 Evaluation Metrics

Accuracy@1 is a metric used in IR to evaluate a system's ability to retrieve a relevant item at the top of the ranking. It measures the proportion of queries for which the most pertinent item appears in the first position. The score ranges from 0 to 1, where 1 indicates perfect retrieval (*i.e.*, the relevant item is consistently ranked first), and 0 means the system never places the appropriate item at the top. This metric is handy when only the top result matters, such as in FAQ matching, question answering, or single-result search scenarios.

While Accuracy@1 is a crucial metric for our primary use case, it only evaluates the top-ranked result. To gain a more comprehensive understanding of retrieval quality, we employed Mean Average Precision at 10 (mAP@10). This metric evaluates the quality of the entire ranked list up to the 10th position, taking into account both the precision and the ranking of relevant items. mAP@10 provides a more nuanced evaluation by rewarding models that place multiple correct items near the top of the list, which is valuable in scenarios where multiple results are used, such as in retrieval-augmented generation pipelines.

[5]Available online at https://huggingface.co/spaces/mteb/leaderboard.

For the Semantic Textual Similarity (STS) evaluation on the *GoSim3* dataset, we used Pearson's correlation coefficient (r). This metric measures the linear correlation between our model's predicted similarity scores (*i.e.*, the cosine similarity of the sentence embeddings) and the ground-truth human judgments. A higher correlation, approaching 1, indicates that the semantic relationships captured by our embeddings strongly align with human perception of similarity, thereby validating the model's ability to discern subtle semantic nuances.

5.2 Baselines

To evaluate the effectiveness of our domain-specific embeddings, we selected pretrained models from the existing literature that have demonstrated superior results in retrieval tasks and sentence representation as baselines. This includes various pretrained models trained using different techniques, encompassing open-source encoders. Additionally, we incorporated a traditional BM-25 model for comparison against the pretrained models. In the following, we summarize these models.

Embeddings from Bidirectional Encoder Representations (E5-models): E5 is a family of advanced text embeddings trained using weakly supervised contrastive Pre-training and a large dataset of text pairs. Our study used the Multilingual E5-base, which is initialized from XLM-RoBERTa weights. The model employs an encoder architecture with average pooling to generate fixed-size embeddings, utilizing cosine similarity for comparison.

BGE M3 is an encoder model designed for multilingual processing and multifunctional tasks. It supports over 100 languages, aiming to streamline text embedding and retrieval for greater efficiency. The model employs self-knowledge distillation, efficient batching, and high-quality data generation to enhance embedding quality. It leverages unsupervised, supervised, and synthesized data through a structured pretraining and fine-tuning approach focused on retrieval tasks.

GTE (Zhang et al., 2024): It refers to a state-of-the-art multilingual encoder specifically designed for retrieval tasks. It was trained using large-scale contrastive learning on a combination of unsupervised, supervised, and synthesized data. This encoder produces dense text embeddings for over 70 languages, ensuring high-quality representations even in long-context scenarios, which is advantageous for industrial applications. Our decision to utilize GTE is based on concepts proposed by an e-commerce company (Alibaba), and it outperforms other models with a similar number of parameters.

Best Matching 25 (BM-25): It is a probabilistic model for IR. It builds on the term frequency (TF) and inverse document frequency (IDF) concepts, such as TF-IDF, but refines term weighting with a non-linear function. This allows BM-25 to rank documents more effectively by considering term frequency and distribution across the corpus, making it better suited for longer documents than TF-IDF.

5.3 Re-Ranking

In addition to the baseline evaluations, we designed an experimental setup where each baseline model is first used to perform semantic search and retrieve the top k candidates most similar to the query. These k candidates, along with the query, are then passed to a re-ranking stage, where a separate model, trained to score semantic similarity, re-evaluates and ranks the candidates to identify the most relevant one. For all experiments, we set $k = 20$. This setup aims to assess the impact of re-ranking within an IR pipeline and determine whether strong encoders alone can eliminate the need for re-ranking.

6 RESULTS

We present the overall results (Subsection 6.1) and then report on our analysis using the Re-ranking approach (Subsection 6.2). Subsection 6.3 presents our dimension effects analysis. Subsection 6.5 presents the ablation study results.

6.1 Overall Findings

Table 1 presents the effectiveness of various models on retrieval datasets evaluated using Accuracy at one. The results include both original and fine-tuned multilingual models, assessed on two datasets: E-FAQ (in Portuguese and Spanish) and *GoSim3*. For the fine-tuned models, we conducted multiple configurations and report those that showed the highest effectiveness on the E-FAQ retrieval task.

Among the domain fine-tuned models, the Multilingual E5 base achieved the highest Accuracy@1 score on the E-FAQ dataset, scoring 90.48% in Portuguese and 90.12% in Spanish. This model performed well on the *GoSim3* dataset, achieving a Pearson Correlation of 43.45%. The fine-tuned XLM model achieved competitive results, with scores of 88.60% in Portuguese and 87.58% in Spanish, yielding the highest Pearson correlation of 48.45% among all models. Similarly, MAP@10 results achieve a

Table 1: Best configuration of finetuned and baseline models' results on retrieval (E-FAQ) and STS (GoSim3) datasets. The E-FAQ scores denote acuraccy@1 (%), and Mean Average Precision MAP@10, and the E-FAQ column corresponds to the test partitions in each considered language. Meanwhile, GoSim3 columns presented the Pearson correlations for Portuguese only.

Group	Model	Embedding Dimension	Parameters (Millions)	E-FAQ				GoSim3
				pt		es		pt
				ACC@1	MAP@10	ACC@1	MAP@10	Pearson
Baseline	*Multilingual E5 Base*	768	279.0	68.98	71.36	70.14	71.36	35.45
	GTE Multilingual		305.0	71.56	74.68	73.90	75.78	35.93
	BGE M3	1024	567.8	73.97	77.27	69.92	73.15	41.05
	BM-25	-	-	76.14	80.27	70.86	73.23	-
Finetuned	*XLM RoBERTa Base*	768	278.0	88.60	90.22	87.58	90.99	48.45
	Multilingual E5 Base		279.0	90.48	92.30	90.12	92.51	43.45

Multilingual E5 base across all tested configurations, yielding 90.48% and 92.51% for Portuguese and Spanish, respectively.

BGE M3 achieved the highest scores over pre-trained models in the E-FAQ evaluation for Portuguese, obtaining 73.97% in Portuguese and 69.92% in Spanish. It also performed best on the STS dataset, obtaining 41.05%. In contrast, the multilingual E5 base model and GTE revealed lower retrieval and STS effectiveness on E-FAQ for Portuguese and *GoSim3*, with accuracy scores of 68.98% and 70.14%, and Pearson coefficients of 35.45% and 35.93%, respectively. Yet, the GTE model surpassed the pre-trained model over E-FAQ in the Spanish partition, achieving 73.90%, followed by the multilingual E5-based model, which registered 70.14%. Related to MAP@10, BGE M3 achieved the highest MAP@10 for Portuguese with 77.27% and 76.15% in Spanish. GTE Multilingual followed with results slightly below, while the original Multilingual E5 Base reached 71.36% in both languages.

The BM-25 baseline outperformed all original pre-trained models on E-FAQ, achieving scores of 76.16% in Portuguese and 70.86% in Spanish. The BM25 baseline yielded MAP@10 scores of 80.27% for Portuguese and 73.23% for Spanish, outperforming some of the pretrained models in Portuguese, but still well below the fine-tuned configurations.

6.2 Reranker Analysis

Figure 4 presents the retrieval effectiveness measured by the Accuracy@1 result for Portuguese across various retrieval models, comparing their results with and without the reranker. For the baseline models (mE5, bge-m3, and gte), applying the reranker generally results in slight improvements or maintains similar accuracy levels. Nevertheless, we observed a minor decrease in performance for BM25 when reranking is applied. The fine-tuned models (F-mE5 and F-XLM) achieved the highest overall accuracy, with both models performing better without reranking—F-mE5 exceeds 90%, while F-XLM reaches nearly 89% Accuracy@1 in the no-reranker setting.

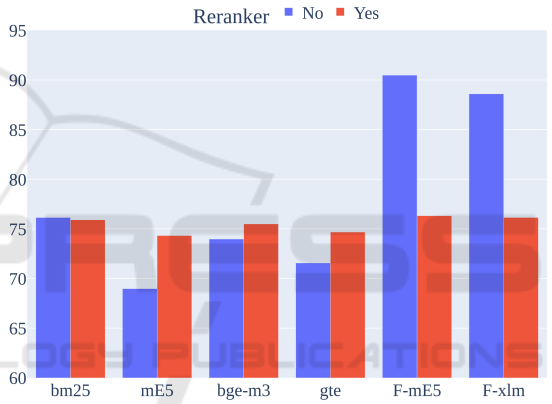

Figure 4: Accuracy at one comparison for **Portuguese** without reranker application for BM25, baseline models, and our best fine-tuned models (F-mE5 and F-xlm).

Figure 5 presents the Accuracy@1 results for Spanish across various retrieval models, comparing configurations with and without reranking. For most baseline models (BM25, mE5, and BGE-M3), applying the reranker yielded slight improvements. We observed a performance drop for GTE when reranking is implemented. The fine-tuned models (F-mE5 and F-XLM) achieved the highest overall accuracy, performing better without reranking. Specifically, F-mE5 achieved approximately 90%, while F-XLM achieved nearly 88% Accuracy@1 without the reranker.

6.3 Dimension Analysis

Figure 6 presents the results of models trained with MLR per the crops embedding dimension from 64 to 768, which affects retrieval effectiveness (Acurracy@1) for the Portuguese test partition of the E-

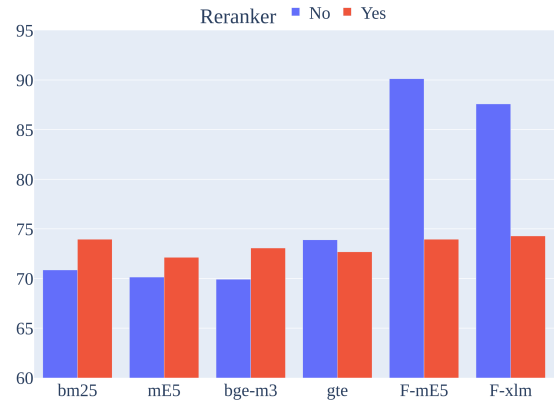

Figure 5: Accuracy at one comparison for **Spanish** without reranker application for BM25, baseline models, and our best fine-tuned models (F-mE5 and F-xlm).

FAQ dataset. All the fine-tuned models (F-mE5 and F-xlm) configurations outperformed the best baseline, BM25, which achieved 76.14%. F-mE5 consistently outperformed F-xlm, with accuracy increasing from 88.07% at dimension 64 to 90.48% at dimension 768. In contrast, F-xlm maintained stable performance, starting at 88.60% and fluctuating to 87.72%. These results indicate that higher dimensions benefit F-mE5 more significantly, while F-xlm is less sensitive to dimensional changes.

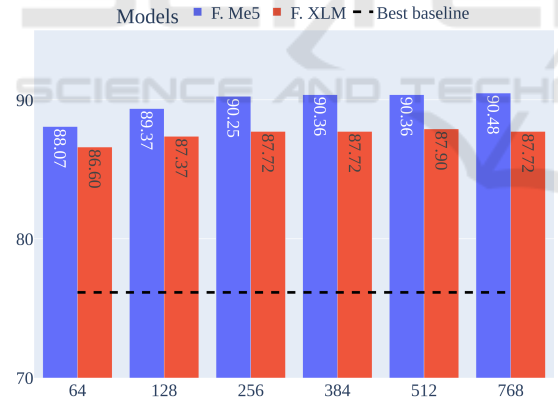

Figure 6: Cropped embedding dimension Accuracy at one value of the trained models on **Portuguese** test partition of E-FAQ; black dashed line represents the best results achieved for BM25 as the best baseline retriever.

We observed similar trends for the Spanish test partition in Figure 7, in which all configurations of the fine-tuned models outperformed the best baseline, GTE multilingual (73.90%). F-mE5 showed a gradual increase in performance with higher embedding dimensions, ranging from 86.87% at dimension 64 to 90.12% at 768. In comparison, F-xlm remained relatively stable, with scores fluctuating slightly between 86.87% and 87.36%. This pattern indicates that higher dimensions might benefit F-mE5 more clearly, while F-xlm seems less influenced by the embedding size.

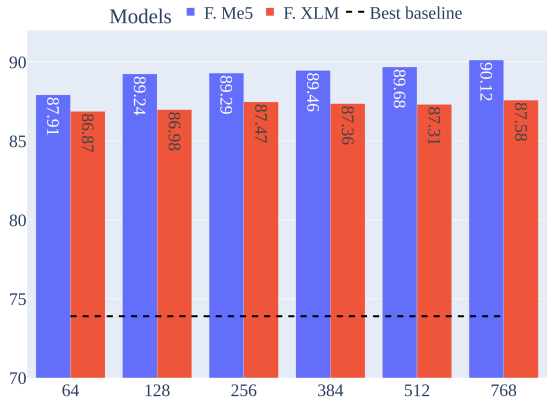

Figure 7: Cropped embedding dimension Accuracy at one value of the trained models on **Spanish** test partition of E-FAQ; black dashed line represents the best results achieved for GTE Multilingual as the best baseline retriever for Spanish.

6.4 Qualitative Results

Table 2 provides qualitative examples of cosine similarity scores for question pairs labeled as similar. The success (S) cases show how the embedding space effectively captures semantic equivalence, even when there are differences in surface forms. In contrast, the failure (F) cases reveal limitations where the cosine score does not match the gold label. These examples underscore both the model's strengths in identifying paraphrases and its weaknesses in addressing nuanced semantic variations.

6.5 Ablation Studies

We investigate the impact of different pretraining methods, the number of hard negatives in contrastive learning, and the combination of loss functions over retrieval and STS benchmarks.

Pretraining Methods. The distinct pretraining approaches of the base models appear to have a direct impact on downstream task outcome. While XLM-RoBERTa relies on a Masked Language Modeling (MLM) objective, the E5 model was pretrained using weakly supervised contrastive learning. Table 1 presents that the E5 model's contrastive foundation provided a significant advantage in our retrieval experiments. This performance gap is statistically significant, confirmed by a comparison of the Average Precision at 10 distributions on the E-FAQ test partition, which yielded a p-value of 7.3×10^{-5} at the Wilcoxon signed-rank test.

Table 2: Qualitative results of cosine similarity scores for question pairs labeled as similar with in their original form and English translated version, illustrating representative Success (S) and Failure (F) cases. The English translation was done by the authors of this work.

Sentences pair	Cosine
cadê a de 25x30 (Where is the 25x30 one?) *Tem como o 25×30* (Is the 25×30 available?)	0.8605 (S)
Vem com módulo ? (Does it come with an amplifier?) *já vem com módulo??* (Does it already include an amplifier?)	0.9689 (S)
Qual a potência de cada saída desse aparelho (What is the output power of each channel of this device?) *Qual a potência do som?* (What is the sound power?)	0.475 (F)
olá, tem em outras cores? (Hello, do you have it in other colors?) *Tem outras cores?* (Are there other colors available?)	0.4793 (F)

Table 3: Qualitative results of cosine similarity scores for question pairs labeled as dissimilar with their respective English translation, illustrating representative Success (S) and Failure (F) cases. The English translation was done by the authors of this work.

Sentences pair	Cosine
Quantos decibéis ele emite? (How many decibels does it emit?) *Qual consumo dele?* (What is its power consumption?)	0.3911 (S)
Cabe no golf mk3 97/98? (Fits golf mk3 97/98) *Boa tarde, tem para Golf 1995.* (Good afternoon, it is available for Golf 1995.)	0.3646 (S)
Bom dia, vocês tem do A51? (Good morning, do you have the A51?) *Bom dia, serve no a51?* (Good morning, does it work on the A51?)	0.6848 (F)
Ja vem com o cooler pro procesador? (Does it come with a CPU cooler?) *Boa noite ja vem com processador?* (Good evening, does it come with a processor?)	0.6773 (F)

Multi-Task Loss. Table 4 presents the results regarding how the multi-task approach contributed to better results on retrieval. The similarity task alone was not sufficient to improve the model's retrieval capacity, as it was unable to determine greater separability on its own. The combination of both tasks yielded the best results for both trained models, as presented in Table 1.

Table 4: Effect of retrieval objective and semantic similarity objective for Multilingual E5 Base on both retrieval (E-FAQ) and STS (GoSim3) datasets. The model was fine-tuned with in-batch negatives only. All columns present the result metrics for Portuguese only data.

Objective	E-FAQ	GoSim3
	ACC@1	Pearson
Similarity only	82.20	54.47
Retrieval only	88.95	39.49
Retrieval & Similarity	90.48	44.57

Hard Negatives. The number of hard negatives extracted from "almost similar" and "dissimilar" labels impacted differently on retrieval and on STS tasks. Table 5 presents this finding. Considering the E5 model trained on retrieval task only, the number of negatives contributes to greater separability and, therefore, a better result on STS, and also increases the quality of retrieval results. With E5 model trained on both retrieval and similarity tasks, we found a optimal value of STS Pearson's correlation using a single hard negative. However, in this scenario, the addition of hard negatives let to a degradation in retrieval accuracy.

Table 5: Effect of hard negatives on InfoNCE loss for Multilingual E5 Base on retrieval (E-FAQ) and STS (GoSim3) datasets. Zero hard negatives indicate in-batch negatives only. All columns present the metrics for Portuguese.

Tasks	Hard Negatives	E-FAQ	GoSim3
		ACC@1	Pearson
Retrieval only	0	88.95	39.49
	1	89.54	46.12
	3	89.13	46.27
Retrieval & Similarity	0	90.48	44.57
	1	89.60	48.03
	3	89.60	46.96

7 DISCUSSION

Table 1 revealed that our fine-tuned, domain-specific models outperformed general sentence encoders on the E-FAQ test set for both Portuguese and Spanish. Even with a domain-specific re-ranking baseline (cf. Figure 4 and 5), our results confirmed the feasibility and effectiveness of using a single, unified embedding model in retrieval pipelines. This key finding corroborates the significant resource optimization potential—reducing memory, CPU processing, and latency—by employing one model instead of two.

Notably, the BM-25 baseline outperformed all original pre-trained models on the E-FAQ dataset.

We attribute this to the inherent characteristics of the e-commerce domain, where related questions frequently contain a significant overlap of specific keywords such as product names, brands, or units of measurement. The effectiveness of our trained sentence encoders suggests that while they grasp the semantic nuances between questions, they also successfully capture this crucial "term-wise" similarity.

Figure 6 demonstrated a favorable trade-off between embedding dimensionality and retrieval effectiveness, underscoring the benefits of MLR training. Our trained models exhibited remarkable effectiveness and stability across various cropped embedding dimensions. Specifically, our top-performing model, F-mE5, achieved a 91.6% reduction in sentence representation size (from 768 to 64 dimensions) while preserving 97.3% of its original retrieval effectiveness.

This dimensionality reduction yields significant practical advantages. Given that most retrieval algorithms scale in memory and time complexity with both the indexed corpus size and the embedding dimension, a 91.6% decrease in embedding size directly correlates to substantial reductions in memory footprint and processing time. Ultimately, this translates to considerably lower demands on computational resources and a more cost-efficient implementation for large-scale retrieval pipelines.

Table 4 and Table 5 underscore the value of our hybrid training methodology. The presented results confirm that a multi-task learning approach achieves a superior balance among retrieval, ranking capabilities, and representation separability. The inclusion of a similarity training task demonstrably enhances both retrieval and semantic textual similarity (STS) results, but only when applied in conjunction with the retrieval task. This improved separability offers practical advantages for semantic retrieval, facilitating the explainability of retrieved elements and enabling the application of similarity score thresholds for result filtering.

While our current investigation focused explicitly on retrieving relevant information within the Portuguese and Spanish e-commerce question paraphrases domain, we are confident that the strengths of our designed multi-objective training methodology offer significant potential for broader generalization. Furthermore, while our study addressed symmetric retrieval for question paraphrases, the adaptability of our models suggests their applicability to a wider range of retrieval tasks, including asymmetric retrieval scenarios, by simply adjusting the training data to a structure similar to, but not restricted to, the E-FAQ.

8 CONCLUSION

Real-world customer inquiries often feature linguistic patterns that challenge traditional sentence encoders and hinder response accuracy. Our study highlighted the effectiveness of domain-specific fine-tuned models for retrieval tasks in Portuguese and Spanish, outperforming the general-purpose pretrained embeddings commonly found in the existing literature. The results demonstrated that our models eliminate the need for additional re-ranking, a process often required when using general embeddings. This makes retrieval more efficient for real-world applications, particularly in E-commerce. Our findings revealed the success of multi-task objective training in Matryoshka Representation Learning, underscoring its relevance in maintaining strong retrieval effectiveness across various embedding dimensions. This is especially advantageous for Portuguese and Spanish, where high-quality retrieval models remain underexplored. Future work will focus on implementing these models in real-world E-commerce environments, with a specific emphasis on the Portuguese and Spanish markets. We will assess their impact on practical real-world applications and refine them for even greater quality in multilingual retrieval. We plan future studies to explore data from other domains or retrieval tasks in a format similar to that proposed for our E-FAQ dataset.

ACKNOWLEDGMENT

This work was supported by *GoBots* company.

REFERENCES

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bednář, J., Náplava, J., Barančíková, P., and Lisický, O. (2024). Some like it small: Czech semantic embedding models for industry applications. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):22734–22742.

Chico, V., Zucchi, L., Ferragut, D., Caus, R., de Freitas, V., and dos Reis, J. C. (2023). Automated question answering via natural language sentence similarity: Achievements for brazilian e-commerce platforms. In *Anais do XIV Simpósio Brasileiro de Tecnologia da*

Informação e da Linguagem Humana, pages 74–83, Porto Alegre, RS, Brasil. SBC.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemi´nski, D., Winata, G. I., Sturua, S., Utpala, S., Ciancone, M., Schaeffer, M., Sequeira, G., Misra, D., Dhakal, S., Rystrøm, J., Solomatin, R., . . . Muennighoff, N. (2025). Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. (2020). CodeBERT: A pre-trained model for programming and natural languages. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547, Online. Association for Computational Linguistics.

Huang, J., Hu, Z., Jing, Z., Gao, M., and Wu, Y. (2024). Piccolo2: General text embedding with multi-task hybrid loss training.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemi´nski, D., Winata, G. I., Sturua, S., Utpala, S., Ciancone, M., Schaeffer, M., Sequeira, G., Misra, D., Dhakal, S., Rystrøm, J., Solomatin, R., . . . Muennighoff, N. (2025). Gemma 3 technical report.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., and Farhadi, A. (2024). Matryoshka representation learning.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Su, J. (2022). Cosent (i): A more effective sentence embedding scheme than sentence-bert. https://kexue.fm/archives/8847. [Online; accessed 12-May-2025].

Tang, Y. and Yang, Y. (2025). Do we need domain-specific embedding models? an empirical investigation.

van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2024a). Text embeddings by weakly-supervised contrastive pre-training.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024b). Multilingual e5 text embeddings: A technical report.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., & Zhang, M. (2024, November). mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In Dernoncourt, F., Preoţiuc-Pietro, D., and Shimorina, A., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.