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

Anonymous ACL submission

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 examines the use of fine-tuned transformer models to generate domain-specific sentence embeddings, specifically for Portuguese and Spanish retrieval tasks. Our findings show that these specialized embeddings significantly outperform general-purpose pretrained models and traditional techniques like BM-25, eliminating the need for additional re-ranking steps in retrieval processes. Our results explore the effects of multi-objective training within Matryoshka Representation Learning, highlighting its effectiveness in maintaining retrieval effectiveness 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 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.

LatinAmericanAI 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 retrieval 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 are releasing our test and calibration datasets. Notably, these datasets are in Portuguese and Spanish, 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[1].

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.

On domain-specific embedding models, Feng et al. (2020) introduced CodeBERT, a transformer-based model trained on open GitHub repositories, which restricts it to six programming languages. It follows multilingual BERT approaches, using mask 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 use the base model to fine-tune the process to improve the alignment between text and code representations. They test the performance of code retrieval based on natural language queries, and CodeBERT outperforms 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 lack of generalization for datasets that differ from the training data.

Regarding language-specific embedding models, 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 6 tasks on CMTEB benchmark, including text retrieval, pair classification, and semantic similarity.

Industrial Applications models (Bednář et al., 2024) focused on creating embedding with 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. The study examines 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 the embedding quality, in addition to measuring search engine ranking effectiveness using precision at 10. Their findings showed that pretrained versions and multilingual distillation provide the best encoder models, highlighting their effectiveness in enhancing search result quality.

DeepFAQ (Chico et al., 2023) is 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 (question and answers). It retrieves candidate questions and

[1]Available at https://anonymous.4open.science/r/Embeddings-Might-Be-All-You-Need-B734/README.md.

applies a domain-specific re-ranking model to identify the most relevant one, ultimately providing the corresponding answer.

Our approach offers a novel and original contribution by utilizing domain-specific embeddings for the e-commerce sector, tailored explicitly 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. (2024) to fine-tune sentence encoding models. These embeddings effectively capture the nuances of informal language used at online platforms, enhancing results in e-commerce-related NLP tasks and addressing gaps identified in previous methods, in particular, the encoder-re-ranker pipeline as presented by 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.

We created this dataset to address a resource gap for Portuguese and Spanish, particularly within the e-commerce domain. We gathered questions from Latin American e-commerce websites sourced from the *LatinAmericanAI* database, as illustrated in Figure 1. 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.

Subsequently, we employed natural language understanding (NLU) models to extract intents and named entities from the questions. This extraction leveraged a machine learning model trained in various entities and intents from existing sentences within the *LatinAmericanAI* data environment. In the NLU context, an intent represents the user's purpose, while an entity represents a term or expression with a known meaning relevant to the sentence's comprehension.

At this point, we had 870,000 sentences in 64 distinct intent categories. Within each category, we employed the HDBSCAN clustering algorithm to group similar questions. This clustering relied on vector representations of the sentences, derived from TF-IDF and singular value decomposition (SVD) applied to the extracted entities. HDBSCAN effectively formed disjoint groups of similar sentences while also identifying and removing noise. This process yielded more than 142,000 clusters, encompassing over 445,000 examples, with the cluster medoid serving as the anchor sentence.

To ensure high-quality semantic similarity data within our clusters, we employed the Gemma 3 language model (Team et al., 2025) to classify each question in the 142,000 clusters as "similar", "almost similar", or "dissimilar" to its anchor sentence. To optimize the classification process, we first curated a separated calibration dataset of 144 real question pairs from e-commerce platforms. Each pair in this dataset was evaluated for similarity by three annotators. The final label for each pair was determined by the majority vote among the annotators. We then identified the specific prompt instructions that yielded the highest accuracy in classifying this calibration dataset according to the majority labels. This best-performing prompt was subsequently used to classify all question pairs within our 142,000 clusters, ensuring a more reliable assessment of semantic similarity. We called the calibration dataset GoSim3, and it is available at HuggingFace's Hub[2].

The dataset was further divided into *training*, *validation*, and *test* sets. The training set comprised most of the data, with 121,248 entries, followed by the validation set, with 13,472 entries. The test sets were divided by language (Portuguese and Spanish) and stratified by intent class, resulting in two sets with 4,000 entries each. We also made the test sets available at HuggingFace's Hub[3].

[2]Available at https://huggingface.co/. The complete URL is omitted for blind review.

[3]Available at https://huggingface.co/. The complete URL is omitted for blind review.

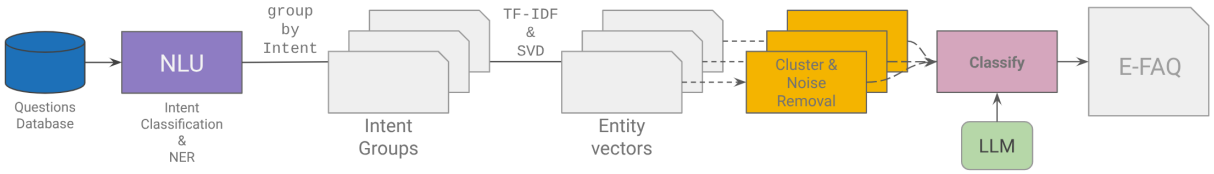

Figure 1: Overview of data collection process.

4 Training Methods

Our proposed models' main application 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). More 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 2 illustrates the architecture. Models that use this architecture, like SBERT (Reimers and Gurevych, 2019), LaBSE (Feng et al., 2022), and E5 (Wang et al., 2024a,b), 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 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 dataset. This method follows Huang et al. (2024) approach.

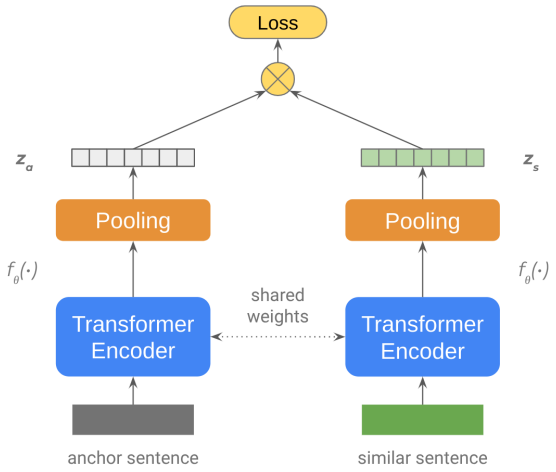

Figure 2: Siamese Dual Encoder model for sentence embeddings generation.

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 $s_{ij} \in \mathbf{S}_i$ a question extracted from the set of questions similar to q_i. The training data consisted in 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.

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:

$$\mathcal{L}_r = \mathcal{L}_{ce} + \mathcal{L}'_{ce} \tag{3}$$

For the semantic similarity objective, we converted the "similar", "almost similar", and "dissimilar" 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:

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

The final multi-task loss is defined by:

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

To achieve our objective of reducing embedding dimensionality, we employed Matryoshka Representation Learning (MLR) (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 at least as accurate as independently trained low-dimensional representations.

For our experiments, we fine-tuned two models: XLM RoBERTa (Conneau et al., 2020), a multilingual transformer model trained with masked language modeling; and the Multilingual E5 Base (Wang et al., 2024b), a model trained for textual representations from RoBERTa. These models generate embeddings with 768 dimensions. We chose the E5 for its favorable ranking on multilingual tasks of the MTEB leaderboard[4].

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

We trained all models in a single GPU, an NVIDIA RTX4090. We trained for at most 5000 training steps with a batch size of 256 pairs of questions. We evaluated the models in every 200 steps, saving the best model checkpoint after validation. We used 2×10^{-5} for learning rate. For the temperature parameter τ we fixed it at 0.05. We trained all models with MLR using $64, 128, 256, 384, 512$, and 768 dimensions.

5 Retrieval Evaluation

This section describes the retrieval evaluation task to assess the quality of our domain-specific embeddings. We outline the evaluation metrics, datasets, and baselines used in this evaluation.

5.1 Evaluation Metric

Accuracy@1 is a metric used in information retrieval 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 zero 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.

5.2 Evaluation Datasets

We selected two datasets for a specific domain but with different purposes: E-FAQ and GoSim3, which validate information retrieval and Semantic Textual Similarity (STS), respectively.

We utilized the test partition from E-FAQ (see Section 3) for a domain-specific dataset, testing both Spanish and Portuguese with 4,000 queries per language.

GoSim3 extends the datasets applied for STS and is oriented to the e-commerce domain. This dataset comprises 144 question pairs labeled as similar, almost similar, and dissimilar. This dataset measures the correlation between human annotations and results obtained by computing the similarity between the vector representations of both questions. In contrast, E-FAQ, this dataset was not used during the model training phase.

5.3 Baselines

To evaluate the effectiveness of our domain-specific embeddings, we selected pretrained models from

the existing literature that have demonstrated superior performance 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. This study used the E5-base, which is initialized from BERT weights. The model utilizes an encoder architecture with average pooling to create fixed-size embeddings, employing 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 pre-training and fine-tuning approach focused on retrieval tasks.

GTE (Zhang et al., 2024): is 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): is a probabilistic model for information retrieval. It builds on term frequency (TF) and inverse document frequency (IDF) concepts like 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.

5.4 Re-ranking

Furthermore, 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 information retrieval pipeline and determine whether strong encoders alone can eliminate the need for re-ranking.

6 Results

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 (Portuguese and Spanish) and GoSim3. Among the domain fine-tuned models, the Multilingual E5 base achieves the highest accuracy at 1 score on the E-FAQ dataset, scoring 90.48% in Portuguese and 90.12% in Spanish. This model also performs well on the GoSim3 dataset, achieving a Pearson Correlation of 0,4345. The fine-tuned XLM model shows competitive results, with scores of 88.60% in Portuguese and 87.58% in Spanish, getting the highest Pearson correlation of 0.4845 over all the models.

BGE M3 achieves the highest scores over pretrained models in the E-FAQ evaluation for Portuguese, obtaining 73.97% in Portuguese and 69.92% in Spanish. It also performs best on the STS dataset, obtaining 0.4105. In contrast, the Multilingual e5 base model and GTE show 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 0.3545 and 0.3593, respectively.

However, the GTE model surpassed the pretrained model over E-FAQ in the Spanish partition, having 73.90%, followed by the multilingual e5-based model, which registered 70.14%.

The BM-25 baseline outperforms all original pretrained models on E-FAQ, achieving scores of 76.16% in Portuguese and 70.86% in Spanish.

Figure 3 presents the retrieval effectiveness measured by the Accuracy@1 result for Portuguese across various retrieval models, comparing their performance with and without the reranker. For the baseline models (mE5, bge-m3, and gte), the application of the reranker generally results in slight

Table 1: Finetuned and baseline models' performances on retrieval datasets (E-FAQ) and STS (GoSim3). The E-FAQ scores denote acuraccy@1 (%), and the E-FAQ column corresponds to the test partitions in each considered language. Meanwhile, GoSim3 columns presented the Pearson correlations for Portuguese only.

	Model	Embedding Dimension	Parameters (Millions)	E-FAQ		GoSim-3
				pt	es	pt
Finetuned	*Multilingual E5 Base*	768	278.0	**90.48**	**90.12**	0.4345
	XLM RoBERTa	768	279.0	88.60	87.58	**0.4845**
Base	*Multilingual E5 Base*	768	278.0	68.98	70.14	0.3545
	GTE Multilingual	768	305.0	71.56	73.90	0.3593
	BGE M3	1024	567.8	73.97	69.92	0.4105
Other	*BM-25*	-	-	76.14	70.86	-

improvements or maintains similar accuracy levels. However, a minor decrease in performance is noted for BM25 when reranking is applied. The fine-tuned models (F-mE5 and F-XLM) achieve 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.

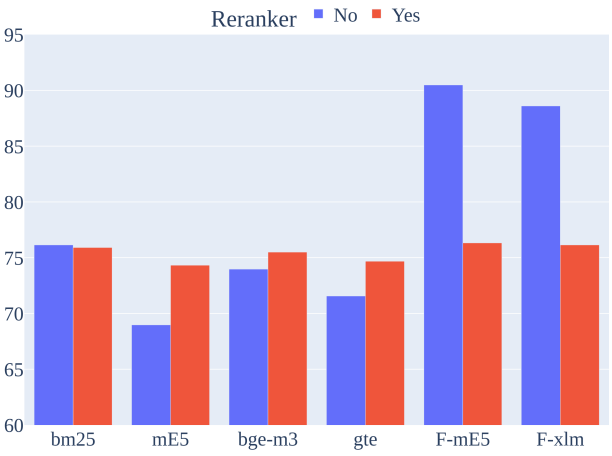

Figure 3: Accuracy at one comparison for **Portuguese** without reranker application for BM25, baseline models, and our best fine-tuned models (F-mE5 and F-xlm).

Figure 4 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 yields slight improvements. We observe 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 reaches approximately 90%, while F-XLM achieves nearly 88% Accuracy@1 without the reranker.

Figure 5 presents the results of the models

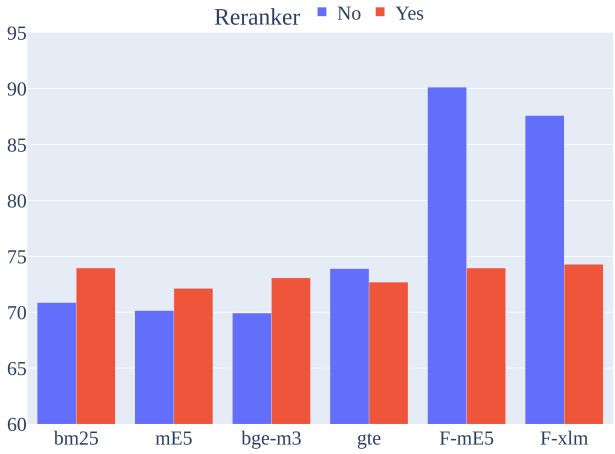

Figure 4: Accuracy at one comparison for **Spanish** without reranker application for BM25, baseline models, and our best fine-tuned models (F-mE5 and F-xlm).

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-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. We observed similar trends for the Spanish results, which are not shown in Figure 5.

7 Discussion

Table 1 revealed that our fine-tuned, domain-specific models outperformed general sentence en-

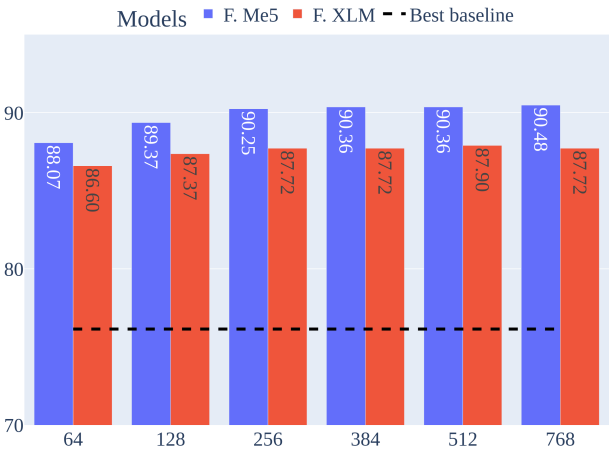

Figure 5: Cropped embedding dimension Accuracy at one value of the trained models on Portuguese test partition of E-FAQ; black dashed line represents the best results achieved for BM25 as the best baseline retriever

coders on the E-FAQ test set for both Portuguese and Spanish. Even with a domain-specific re-ranking baseline (cf. Figure 3 and 4), 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 performed better than 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 5 illustrates a favorable trade-off between embedding dimensionality and retrieval effectiveness, underscoring the benefits of MLR training. Our trained models exhibit remarkable effectiveness and stability across various cropped embedding dimensions. Specifically, our top-performing model, F-mE5, achieves 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.

While our current investigation focused explicitly on retrieving relevant information within the Portuguese e-commerce question paraphrases domain, we are confident that the strengths of the designed multi-objective training methodology offer significant potential for broader generalization. Furthermore, while our study addresses symmetric retrieval for question paraphrases, the adaptability of our models suggests their applicability to a broader range of retrieval tasks, including asymmetric retrieval scenarios, simply by adjusting the training data accordingly 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 multi-task objective training success in Matryoshka Representation Learning by 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 emphasize implementing these models in real-world E-commerce environments, specifically for 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. Future studies can also explore data from other domains or retrieval tasks in a format similar to that proposed for the E-FAQ.

Limitations

This section highlights potential threats to the quality of our research study, focusing on three categories (Petersen and Gencel, 2013): internal validity, external validity, and conclusion validity.

Internal Validity: Our experimental reliability is directly tied to dataset quality. Despite efforts to improve it with AI-generated pseudo-labels, potential biases or imbalances in the data could still impact our results and might not entirely reflect real-world conditions.

External Validity: Our study is limited to the e-commerce domain and symmetric retrieval settings. While this allows for controlled experimentation, we should be cautious about generalizing our findings to other domains. Future research should explore their applicability in contexts like customer support systems.

Conclusion Validity: We must use more robust hypothesis testing methods to ensure our findings are statistically significant. Current methods may not adequately account for data variability, so more statistical tests will help distinguish meaningful patterns from random noise.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Víctor Chico, Luiz Zucchi, Daniel Ferragut, Rodrigo Caus, Victor de Freitas, and Julio Cesar dos Reis. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *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.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. CodeBERT: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547, Online. Association for Computational Linguistics.

Junqin Huang, Zhongjie Hu, Zihao Jing, Mengya Gao, and Yichao Wu. 2024. Piccolo2: General text embedding with multi-task hybrid loss training. *Preprint*, arXiv:2405.06932.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka representation learning. *Preprint*, arXiv:2205.13147.

Kai Petersen and Cigdem Gencel. 2013. Worldviews, research methods, and their relationship to validity in empirical software engineering research. In *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, pages 81–89.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Yixuan Tang and Yi Yang. 2025. Do we need domain-specific embedding models? an empirical investigation. *Preprint*, arXiv:2409.18511.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.