

# Efficient Italian-Specific Embedding Model using Contrastive Fine-Tuning

Archit Rastogi  
Matricola: 1982785

## 1. Motivation

The [MTEB leaderboard](#) is a comprehensive benchmark platform that ranks embedding models across a wide range of languages and tasks. In this project, we focus specifically on the retrieval category, zero-shot evaluations, and open-source models relevant to Italian language performance.

Current top-performing Italian retrieval models on MTEB are dominated by very large multilingual models such as [Qwen3-Embedding-4B](#) and [Qwen3-Embedding-8B](#), which achieve scores between 86–90. While impressive, these models have billions of parameters, are resource-intensive, and not tailored for Italian-specific semantics.

Smaller multilingual models like [multilingual-e5-small](#) (118M parameters) are easier to deploy but reach only around 77 points on Italian retrieval benchmarks, leaving a 9–13 point gap.

This project explores whether language-specific fine-tuning can bridge this gap by leveraging Italian-only data and contrastive learning to create a smaller yet high-performing embedding model.

## 2. Objective

The goal is to develop a compact ( $< 200\text{M}$ ) embedding model that:

- Outperforms [multilingual-e5-small](#) by at least 5 points on Italian tasks.
- Is specifically optimized for Italian text retrieval and semantic understanding.
- Maintains low computational and memory cost for real-world deployment.
- Produces a high-quality, publicly available Italian retrieval fine-tuning dataset to support future research and model development in this underserved language.

## 3. Approach

**Base Model:** [multilingual-e5-small](#) serves as the foundation, with 118M parameters and 384-dimensional embeddings. It is a proven model for multilingual semantic retrieval and provides a stable training baseline.

**Training Data:** A curated Italian retrieval dataset will be constructed from the Italian subset of the

[Google C4 dataset](#). The data will be transformed into query–document pairs. This dataset will be made publicly available on HuggingFace to enable reproducibility and support the Italian NLP research community, which currently lacks dedicated resources for retrieval fine-tuning.

**Evaluation Data:** Performance will be tested on the Italian benchmarks from [BelebeleRetrieval](#) (Italian subset) and [WikipediaRetrievalMultilingual](#). These datasets capture both monolingual and multilingual retrieval capabilities.

**Training Strategy:** The fine-tuning will use contrastive learning with the InfoNCE loss. Positive pairs are Italian sentence–document pairs, while in-batch negatives serve as contrastive samples.

## 4. Evaluation

Performance will be evaluated using standard retrieval metrics NDCG@10 and Recall@10 on both Italian datasets. The aim is to achieve a meaningful gain over the multilingual baseline while keeping inference cost low.

We will also compare the proposed model with larger multilingual models, such as [Qwen3-Embedding-4B](#), as well as their quantized versions. This will help create a fair and unbiased evaluation, allowing us to analyze the trade-offs between computational efficiency and retrieval accuracy across model scales.

## 5. Expected Outcomes

This project will deliver a small, efficient embedding model specialized for Italian retrieval tasks. Expected contributions include:

- Demonstrating that focused Italian-specific fine-tuning can significantly improve retrieval accuracy for small models.
- Providing a deployable Italian embedding model suitable for search, recommendation, and NLP pipelines.
- Releasing a high-quality, publicly accessible Italian retrieval fine-tuning dataset on HuggingFace, addressing a significant gap in available resources for Italian language model development.
- Establishing a scalable framework for extending this approach to other underrepresented languages.