Italian Retrieval Embeddings: When MTEB Scores Mislead

December 21, 2025

Archit Rastogi

Abstract

We expose a critical flaw in MTEB leaderboard evaluation: cross-language score aggregation causes practitioners to systematically underestimate small multilingual models. For Italian retrieval, `multilingual-e5-small` (118M params) achieves 90.8 NDCG@10—only 2.9 points below models with 560M parameters—despite showing 77 on MTEB's aggregated score. This 13.8-point discrepancy has significant deployment implications. We release a 25M Italian triplet dataset and demonstrate that fine-tuning already-saturated models causes catastrophic forgetting (11–52% drops). Code and data: https://huggingface.co/datasets/ArchitRastogi/it-retrieval-triplets-mc4

1. Introduction

Main Contribution. We expose a fundamental measurement issue in the MTEB leaderboard (Muennighoff et al., 2022): cross-language score aggregation masks language-specific performance, causing practitioners to systematically underestimate small multilingual models. When filtering MTEB for Italian retrieval, `multilingual-e5-small` (118M params) displays an overall score of 77.0, seemingly requiring 13+ points of improvement to match larger models. However, when evaluated only on Italian tasks, this model achieves **90.8 NDCG@10**—only 2.9 points below `multilingual-e5-large` (560M params) and 4.6 points below Qwen3-8B.

Deployment Impact. This measurement flaw has real consequences: practitioners deploy 8B parameter models (requiring ∼16GB VRAM) when 118M models (requiring ∼500MB) achieve 90.8 vs 95.4 NDCG@10—a **68× reduction in model size** for a 4.6-point performance gap.

Email: Archit Rastogi <rastogi.1982785@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

Secondary Contributions. To investigate whether language-specific fine-tuning could bridge the perceived gap, we created a 25M Italian retrieval triplet dataset from mC4 (Raffel et al., 2020) and a custom 1,000-query RAG benchmark following RAGEval (Zhu et al., 2024). Fine-tuning experiments revealed catastrophic forgetting (11–52% drops), providing evidence that small multilingual models already saturate Italian retrieval performance.

2. Related Work

E5 (Wang et al., 2022) and BGE-M3 (Chen et al., 2024) are leading multilingual embedding models trained with contrastive learning. Both achieve strong MTEB performance, but leaderboard scores aggregate across all language subsets, obscuring language-specific capabilities. InfoNCE loss with hard negatives is effective for dense retrieval (Karpukhin et al., 2020). Catastrophic forgetting (Kirkpatrick et al., 2017) occurs when neural networks trained on new data forget previously learned information. No prior work has systematically analyzed MTEB's aggregation effects on multilingual model evaluation or created large-scale Italian retrieval datasets.

3. Method

3.1. Baselines: MTEB Italian Evaluation

We evaluated 22 multilingual embedding models on Italian-only MTEB tasks: BelebeleRetrieval (cross-lingual QA) and WikipediaRetrievalMultilingual (encyclopedic retrieval). For each model, we computed: (1) overall MTEB score (aggregated across all languages), and (2) Italian-only average (mean of Belebele Italian and Wikipedia Italian monolingual NDCG@10). This reveals the gap between global aggregation and language-specific performance. BM25 serves as a lexical baseline.

3.2. Contribution: Dataset and Benchmark

Italian Retrieval Triplets (25M). We constructed triplets from mC4 Italian (Raffel et al., 2020) following the query-positive-negative paradigm. Queries are first sentences of paragraphs (mean 17 tokens); positives are

Table 1. MTEB Italian retrieval (NDCG@10). Overall score vs. Italian-only reveals 13.8-point discrepancy for `ml-e5-small`.

Model	MTEB	Bele.	Wiki	Avg
Qwen3-Emb-8B	90.4	98.7	92.1	95.4
Qwen3-Emb-4B	86.2	95.3	89.8	92.5
ml-e5-large	84.2	95.1	92.3	93.7
BGE-M3	84.0	93.7	90.2	91.9
ml-e5-small	77.0	92.4	89.3	**90.8**
e5-small-v2	41.6	74.1	77.7	75.9
After fine-tuning on 25M Italian triplets:				
ml-e5-small (ft)	77.0	84.2	82.1	83.1
e5-small-v2 (ft)	41.6	37.3	41.9	39.6

remaining paragraph text (mean 47 tokens). Hard negatives (8.6M) were retrieved via `multilingual-e5-small` from length-matched buckets; random negatives (16.8M) were randomly sampled. Total: 25.4M triplets with zero exact duplicates. Dataset published on HuggingFace.

Custom RAG Benchmark. Following RAGEval methodology (Zhu et al., 2024), we created an Italian evaluation set with 1,000 queries, 216 documents, and 6,885 chunks from legal/administrative Italian text to test retrieval on domain-specific content distinct from Wikipedia/Belebele.

Fine-tuning Setup. We fine-tuned `multilingual-e5-small` and `e5-small-v2` using MultipleNegativesRankingLoss (τ=0.05), AdamW optimizer (lr $\in$ {5e-6, 1.5e-5}), batch sizes 192–384, on NVIDIA RTX 3090. Optuna hyperparameter search (8 trials) explored learning rate, batch size, temperature, and warmup. Fine-tuned models available on HuggingFace: ml-e5-small-ft and e5-small-v2-ft.

4. Experiments

4.1. MTEB Aggregation Reveals Hidden Performance

Table 1 demonstrates the core finding: `ml-e5-small` achieves 90.8 NDCG@10 on Italian tasks—only 2.9 points below `ml-e5-large` (560M params, 93.7 NDCG)—despite showing 77.0 vs 84.2 on aggregated MTEB scores (7-point gap). This 13.8-point discrepancy between overall and Italian-only scores systematically undervalues small models. The gap narrows for larger models: Qwen3-8B shows only 5.0 points difference (90.4 overall vs 95.4 Italian). See Appendix Figure 1 for visualization.

4.2. Fine-tuning Causes Catastrophic Forgetting

Fine-tuning on 25M Italian triplets consistently degraded performance across all hyperparameter configurations, providing evidence of task saturation (see Appendix Table 2 for full results): `ml-e5-small`: 0.488 $\rightarrow$ 0.434 NDCG (-11%); `e5-small-v2`: 0.408 $\rightarrow$ 0.209 NDCG (-49%). All 8 Optuna trials showed degradation, confirming this indicates fundamental saturation rather than configuration issues. Both fine-tuned models perform worse than BM25 (0.361 NDCG), despite base models significantly outperforming it. The severe collapse of `e5-small-v2` demonstrates how hard negative mining with the target model creates circular training dynamics that amplify forgetting. On MTEB Italian tasks, fine-tuned `e5-small-v2` dropped from 75.9 to 39.6 average NDCG@10 (-48%).

5. Conclusion

We demonstrate that MTEB's cross-language aggregation masks language-specific performance, leading practitioners to underestimate small multilingual models. For Italian retrieval, `multilingual-e5-small` (118M params) achieves 90.8 NDCG@10—only 2.9 points below 560M parameter models—despite a 7-point overall MTEB gap. Fine-tuning such saturated models causes catastrophic forgetting (11–52% drops), even with extensive hyperparameter tuning. Key takeaways: (1) always evaluate multilingual models on target language subsets rather than aggregated scores; (2) for saturated tasks (base NDCG $>$ 0.90), language-specific fine-tuning degrades performance; (3) hard negative mining with the target model amplifies forgetting. We release the first large-scale Italian retrieval dataset to support future research.

References

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pp. 6769–6781, 2020.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Zhu, K., Luo, Y., Xu, D., Wang, R., Yu, S., Wang, S., Yan, Y., Liu, Z., Han, X., Liu, Z., and Sun, M. Rageval: Scenario specific rag evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262*, 2024.