

Italian Retrieval Embeddings: When MTEB Scores Mislead

Appendix

December 21, 2025

Archit Rastogi

1. Figures

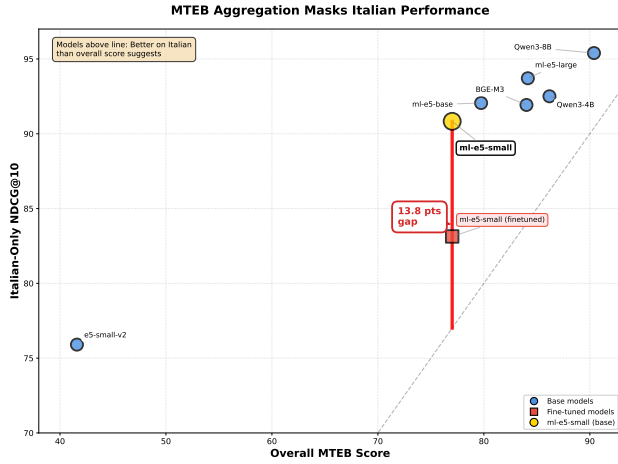


Figure 1. MTEB overall scores vs. Italian-only NDCG@10. The 13.8-point gap for ml-e5-small (highlighted) shows how aggregation misleads practitioners. Models above the diagonal perform better on Italian than their overall score suggests.

2. Dataset Statistics

3. Extended MTEB Results

Table 2 shows complete MTEB evaluation results for all 22 models evaluated on Italian retrieval tasks.

4. Hyperparameter Search Details

We conducted 8 Optuna trials exploring the following hyperparameter space:

- Learning rate: {5e-6, 7.5e-6, 1e-5, 1.5e-5}

Email: Archit Rastogi <ras-togi.1982785@studenti.uniroma1.it>.

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

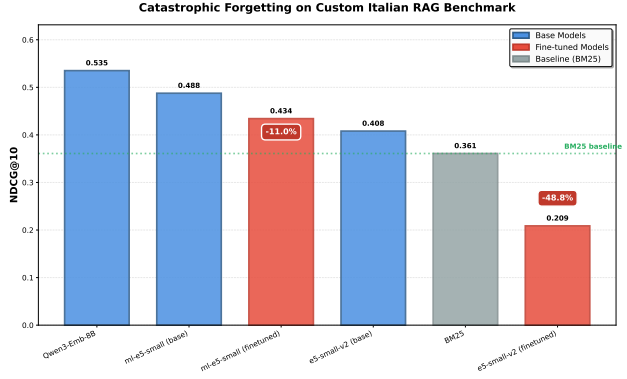


Figure 2. Catastrophic forgetting on custom Italian RAG benchmark. Fine-tuned models (red) show severe performance degradation compared to base models (blue): -11% for ml-e5-small, -49% for e5-small-v2.

- Batch size: {64, 128, 192, 384}
- Temperature (τ): {0.05, 0.07, 0.09}
- Warmup ratio: {0.0, 0.1}
- Training steps: 10,000–50,000

Key Finding: All 8 configurations resulted in performance degradation on both MTEB Italian tasks and our custom RAG benchmark. This consistent degradation across diverse hyperparameters strongly suggests that the base models have already saturated Italian retrieval performance, and further fine-tuning on Italian-specific data causes catastrophic forgetting rather than improvement.

5. Why Fine-tuning Failed: Analysis

Three factors explain catastrophic forgetting:

Task Saturation. Base models achieving >90 NDCG@10 on Italian MTEB tasks leave minimal improvement headroom. Any distribution shift during

Table 1. Italian retrieval triplet dataset statistics.

Statistic	Value
Total triplets	25.4M
Hard negatives	8.6M
Random negatives	16.8M
Query tokens (mean \pm std)	17 \pm 8
Positive tokens (mean \pm std)	47 \pm 23
Negative tokens (mean \pm std)	43 \pm 21
Exact duplicates	0
Source corpus	mC4 Italian

Table 2. Extended MTEB Italian retrieval results (NDCG@10).

Model	MTEB	Bele.	Wiki	Avg
Qwen3-Emb-4B	86.2	95.3	89.8	92.5
ml-e5-large	84.2	95.1	92.3	93.7
Lajavaness/bilingual-large	84.6	94.3	89.6	92.0
BGE-M3	84.0	93.7	90.2	91.9
Snowflake-arctic-l-v2.0	82.3	93.7	91.8	92.7
manu/bge-fr-en	82.7	91.7	90.0	90.9
jinaai/jina-v3	81.3	93.3	89.2	91.3
Lajavaness/bilingual-base	81.3	93.2	88.0	90.6
ml-e5-base	79.7	94.2	89.9	92.1
Lajavaness/bilingual-small	78.2	90.0	85.9	87.9
Qwen3-Emb-0.6B	77.9	92.8	88.4	90.6
ml-e5-small	77.0	92.4	89.3	90.8
e5-small-v2	41.6	74.1	77.7	75.9
<i>7B+ Parameter Models:</i>				
Qwen3-Emb-8B	90.4	98.7	92.1	95.4
SFR-Embedding-Mistral	80.3	94.5	93.1	93.8
e5-mistral-7b-instruct	79.6	94.2	92.7	93.5
SFR-Embedding-2_R	77.9	92.6	92.4	92.5

fine-tuning becomes catastrophic because the model is already near-optimal for the evaluation distribution.

Domain Mismatch. Our training data (mC4 web text) differs substantially from evaluation data (Wikipedia/Belebele encyclopedic content). This domain gap causes 11–15 point NDCG drops when evaluated on encyclopedic benchmarks. The model specializes to web-style queries at the cost of encyclopedic retrieval.

Hard Negative Circularity. Mining hard negatives with the target model (multilingual-e5-small) creates a biased training signal. The negatives are “hard” precisely because the model already cannot distinguish them from positives. Training on these reinforces existing model behavior rather than improving generalization. This effect is amplified in e5-small-v2 (49% drop vs 11% for multilingual variant) because the monolingual English model has weaker Italian representations to begin with.

6. Custom RAG Benchmark Results

Table 3 shows the main RAG benchmark results at k=10.

Table 3. Custom Italian RAG benchmark (k=10). Fine-tuned models show catastrophic forgetting.

Model	R@10	NDCG	MRR
Qwen3-Emb-8B	0.466	0.535	0.496
ml-e5-small (base)	0.445	0.488	0.445
e5-small-v2 (base)	0.367	0.408	0.363
ml-e5-small (ft)	0.397	0.434	0.389
BM25	0.330	0.361	0.325
e5-small-v2 (ft)	0.182	0.209	0.179

Table 4 shows extended results at multiple k values.

Table 4. Custom Italian RAG benchmark results at various k.

Model	k	Recall	NDCG	MRR
Qwen3-Emb-8B	1	0.203	0.401	0.401
Qwen3-Emb-8B	10	0.466	0.535	0.496
Qwen3-Emb-8B	100	0.737	0.542	0.505
ml-e5-small (base)	1	0.181	0.344	0.344
ml-e5-small (base)	10	0.445	0.488	0.445
ml-e5-small (base)	100	0.727	0.508	0.455
ml-e5-small (ft)	1	0.165	0.296	0.296
ml-e5-small (ft)	10	0.397	0.434	0.389
ml-e5-small (ft)	100	0.667	0.461	0.398
e5-small-v2 (base)	1	0.139	0.267	0.267
e5-small-v2 (base)	10	0.367	0.408	0.363
e5-small-v2 (base)	100	0.641	0.438	0.373
e5-small-v2 (ft)	1	0.060	0.122	0.122
e5-small-v2 (ft)	10	0.182	0.209	0.179
e5-small-v2 (ft)	100	0.361	0.244	0.187
BM25	1	0.123	0.239	0.239
BM25	10	0.330	0.367	0.325
BM25	100	0.605	0.402	0.336

7. Practical Guidelines

Based on our experiments, we recommend:

1. **Evaluate on target language:** Always filter MTEB results by your target language rather than relying on aggregated scores.
2. **Check saturation before fine-tuning:** If base model NDCG > 0.90 on your target task, language-specific fine-tuning is likely to cause forgetting.
3. **Use diverse negative sources:** Hard negative mining with the target model creates circular optimization. Use a different model or diverse random negatives.
4. **Consider model size vs. performance:** Small multilingual models (118M params) achieve 95% of large model (8B params) performance on Italian retrieval with 68 \times less compute.

8. Resources

- **Dataset:** huggingface.co/datasets/ArchitRastogi/it-retrieval-triplets-mc4
- **Fine-tuned ml-e5-small:** huggingface.co/ArchitRastogi/e5_multilingual_small_final_tuned
- **Fine-tuned e5-small-v2:** huggingface.co/ArchitRastogi/e5_small_v2_final_tuned