Italian Retrieval Embeddings: When MTEB Scores Mislead

Archit Rastogi
Sapienza University of Rome
Department of Computer Science
`rastogi.1982785@studenti.uniroma1.it`

Abstract

We identify a systematic limitation in MTEB leaderboard evaluation: aggregating scores across languages can obscure language-specific performance and undervalue smaller multilingual models in practical deployment settings. Focusing on Italian retrieval, `multilingual-e5-small` (118M parameters) achieves 90.8 NDCG@10 on Italian-only tasks only 2.9 points below models with 560M parameters despite scoring 77.0 on MTEB's aggregated leaderboard, a discrepancy of 13.8 points. This mismatch has concrete deployment implications: practitioners may select 8B-parameter models requiring ∼16GB VRAM, even though a 118M-parameter model requiring ∼500MB attains comparable performance on the target language. To examine whether language-specific fine-tuning can close this apparent gap, we release a 25.4M Italian retrieval triplet dataset and a 1,000-query Italian RAG benchmark. Across extensive fine-tuning experiments, we observe consistent performance degradation (11–49%), indicating saturation of Italian retrieval performance in small multilingual models under this setting. Code and data are available at `https://github.com/ArchitRastogi20/italian-retrieval-mteb`.

1 Introduction

The Massive Text Embedding Benchmark (MTEB; Enevoldsen et al., 2025) has become the de facto standard for evaluating text embedding models, with practitioners routinely selecting models based on leaderboard rankings. However, MTEB's cross-language score aggregation can obscure language-specific performance differences.

Main Contribution. We expose a systematic evaluation artifact in MTEB evaluation: when filtering for Italian retrieval tasks specifically, `multilingual-e5-small` (Wang et al., 2024) (118M parameters; hereafter `ml-e5-small`) achieves 90.8 NDCG@10—only 2.9 points below `multilingual-e5-large` (560M parameters; hereafter `ml-e5-large`) at 93.7 and 4.6 points below `Qwen3-Embedding-8B` (Zhang et al., 2025) (hereafter `Qwen3-8B`) at 95.4. Yet on MTEB's aggregated leaderboard, these models score 77.0, 84.2, and 90.4 respectively, creating a misleading 13+ point gap that does not reflect actual Italian performance. We emphasize that our analysis focuses exclusively on Italian retrieval tasks within MTEB, and our conclusions should be interpreted within this scope.

Deployment Impact. This measurement flaw has real consequences. Practitioners see the MTEB leaderboard and may conclude they need an 8B parameter model (requiring ∼16GB VRAM, embedding dimension 4096) when a 118M model (requiring ∼500MB, embedding dimension 384) achieves 95% of the performance on their target language a 68× reduction in model size for only 4.6 points of NDCG.

Secondary Contributions. To investigate whether language-specific fine-tuning could bridge the perceived gap, we created:

1. A **25.4M Italian retrieval triplet dataset** from mC4 (Raffel et al., 2020) with both hard and random negatives.
2. A **custom 1,000-query RAG benchmark** following RAGEval methodology (Zhu et al., 2024) on Italian legal/administrative text.

Fine-tuning experiments revealed catastrophic forgetting (11–49% drops across models), providing evidence that small multilingual models have already saturated Italian retrieval performance, and further language-specific training degrades rather than improves results.

2 Related Work

Multilingual Embedding Models. E5 (Wang et al., 2024) and BGE-M3 (Chen et al., 2024) represent leading approaches to multilingual text embeddings, trained with contrastive learning on large-scale multilingual corpora. While both achieve strong MTEB performance, the leaderboard aggregates scores across all language subsets, potentially obscuring language-specific capabilities.

Contrastive Learning for Retrieval. InfoNCE loss with hard negative mining has proven effective for dense retrieval (Karpukhin et al., 2020). The quality of negative samples significantly impacts learned representations, with hard negatives mined from the embedding space typically outperforming random negatives.

Catastrophic Forgetting. Neural networks trained on new data often forget previously learned information (Kirkpatrick et al., 2017). This phenomenon is particularly relevant for fine-tuning pretrained models, where domain-specific training can degrade general capabilities.

Gap in Prior Work. No prior work has systematically analyzed MTEB's aggregation effects on multilingual model evaluation for specific languages, nor created large-scale Italian-specific retrieval datasets to investigate this phenomenon.

3 Method

3.1 MTEB Italian Evaluation

We evaluated 22 multilingual embedding models on Italian-only MTEB tasks to quantify the gap between aggregated and language-specific performance.We report the official MTEB aggregated score as provided by the public leaderboard.

Tasks. We focus on two Italian retrieval benchmarks available in MTEB:

- **BelebeleRetrieval**: Cross-lingual question answering retrieval with Italian queries and passages.
- **WikipediaRetrievalMultilingual**: Encyclopedic retrieval on Italian Wikipedia, including both monolingual (Italian-to-Italian) and cross-lingual (Italian-to-English, English-to-Italian) settings.

These tasks were automatically selected by filtering the MTEB leaderboard for Italian language tasks.

Metrics. For each model, we compute:

1. **Overall MTEB score**: The aggregated score across all languages as reported on the MTEB leaderboard.
2. **Italian-only average**: Mean NDCG@10 of BelebeleRetrieval (Italian) and WikipediaRetrievalMultilingual (Italian monolingual).

This comparison reveals discrepancies between global rankings and language-specific performance.

Baseline. BM25 serves as a lexical baseline to contextualize neural model improvements.

3.2 Italian Retrieval Triplet Dataset

We constructed a large-scale Italian retrieval dataset to enable fine-tuning experiments.

Source Corpus. We extracted text from mC4 Italian (Raffel et al., 2020), the Italian subset of the multilingual C4 corpus derived from Common Crawl.

Triplet Construction. Following the query-positive-negative paradigm:

- **Queries**: First sentences of paragraphs (mean 17 tokens, std 8).
- **Positives**: Remaining paragraph text (mean 47 tokens, std 23).
- **Negatives**: A mixture of hard and random negatives.

Hard Negative Mining. Hard negatives (8.6M) were retrieved using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019). For each query, we:

1. Encoded all documents in length-matched buckets.
2. Retrieved the top-10 most similar documents (excluding the positive).
3. Randomly selected one as the hard negative.

Length-matched buckets ensure negatives have similar surface characteristics to positives, forcing the model to learn semantic distinctions rather than length heuristics.

Random Negatives. Random negatives (16.8M) were sampled uniformly from the corpus, providing diversity in the training signal.

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

Table 1: Italian retrieval triplet dataset statistics.

Dataset Statistics. The final dataset contains 25.4M triplets with zero exact duplicates. Table 1 summarizes the statistics.

3.3 Custom RAG Benchmark

To evaluate models on domain-specific Italian retrieval beyond MTEB's encyclopedic content, we created a custom benchmark following RAGEval methodology.

Construction. We translated passages from existing RAGEval datasets to Italian using GPT-4o-mini (OpenAI, 2024), then generated Italian questions and ground-truth relevance labels using the same model. This approach avoids the pitfalls of simple translation, which can break query-document alignment. We use GPT-4o-mini to preserve semantic alignment between queries and documents, as direct translation can distort retrieval relevance.

Statistics. The benchmark contains 1,000 queries, 216 documents, and 6,885 chunks from Italian legal and administrative text, providing a challenging domain-specific evaluation distinct from Wikipedia-style content.

3.4 Fine-tuning Setup

We fine-tuned two models to investigate whether Italian-specific training improves performance:

Models.

- `ml-e5-small` (118M parameters): Strong multilingual baseline.
- `e5-small-v2` (33M parameters): English-centric model for comparison. (Wang et al., 2022)

Training Configuration.

- **Loss**: MultipleNegativesRankingLoss with temperature $\tau = 0.05$.

Model	**MTEB$_{\text{agg}}$**	**Italian-only (NDCG@10)**		
		Bele.	**Wiki**	**Avg**
Qwen3-8B	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 (Italian-only eval.):				
ml-e5-small (ft)	–	84.2	82.1	83.1
e5-small-v2 (ft)	–	37.3	41.9	39.6

Table 2: Italian retrieval results on MTEB tasks (NDCG@10). **MTEB$_{\text{agg}}$** is the official aggregated score across all languages; Italian-only scores average BelebeleRetrieval (Italian) and WikipediaRetrievalMultilingual (Italian).

- **Optimizer**: AdamW with learning rates in {5e-6, 1.5e-5}.
- **Batch sizes**: 192–384 with gradient accumulation.
- **Hardware**: NVIDIA RTX 3090 (24GB).

Hyperparameter Search. We conducted 8 Optuna trials exploring learning rate, batch size, temperature, and warmup ratio. All configurations were evaluated on both MTEB Italian tasks and our custom RAG benchmark.

4 Experiments

4.1 MTEB Aggregation Masks Italian Performance

Table 2 shows that the largest discrepancy between aggregated and Italian-only performance occurs for smaller multilingual models.

Key Observations. (1) Aggregated scores obscure small models. `ml-e5-small` exhibits a 13.8-point discrepancy between its aggregated MTEB score (77.0) and its Italian-only average (90.8). This discrepancy is substantially smaller for larger models, with `Qwen3-8B` showing a 5.0-point difference (90.4 vs. 95.4).

(2) Near-equivalent performance at different scales. On Italian tasks, `ml-e5-small` (118M parameters, 384 dimensions) attains 90.8 NDCG@10, compared to 93.7 for `ml-e5-large` (560M parameters, 1024 dimensions), a difference of only 2.9 points despite a 4.7× increase in model size.

Model	R@10	NDCG	MRR
Qwen3-8B	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 3: Custom Italian RAG benchmark results at $k = 10$. Fine-tuned models (ft) perform worse than their base versions.

(3) Diminishing returns beyond mid-scale models. Scaling from `ml-e5-large` (560M) to `Qwen3-8B` (8B) yields a further gain of 1.7 NDCG@10 points (93.7 → 95.4) while increasing model size by approximately 14×.

4.2 Fine-tuning Causes Catastrophic Forgetting

All fine-tuning configurations degraded performance, providing evidence that base models have saturated Italian retrieval capabilities.

MTEB Results. Fine-tuned models show substantial drops on MTEB Italian tasks:

- `ml-e5-small`: 90.8 → 83.1 (−8.5%)
- `e5-small-v2`: 75.9 → 39.6 (−48%)

Custom RAG Benchmark. Table 3 shows results on our domain-specific benchmark at $k = 10$.

Forgetting Severity.

- `ml-e5-small`: 0.488 → 0.434 NDCG (−11%)
- `e5-small-v2`: 0.408 → 0.209 NDCG (−49%)

The severe collapse of `e5-small-v2` (below BM25 after fine-tuning) demonstrates how English-centric models with weaker Italian representations are more susceptible to catastrophic forgetting.

4.3 Analysis: Why Fine-tuning Failed

Three factors explain the consistent performance degradation:

Task Saturation. Base models achieving >90 NDCG@10 on MTEB Italian tasks leave minimal improvement headroom. At this performance level, the model is already near-optimal for the evaluation distribution, and any distribution shift during fine-tuning risks degradation.

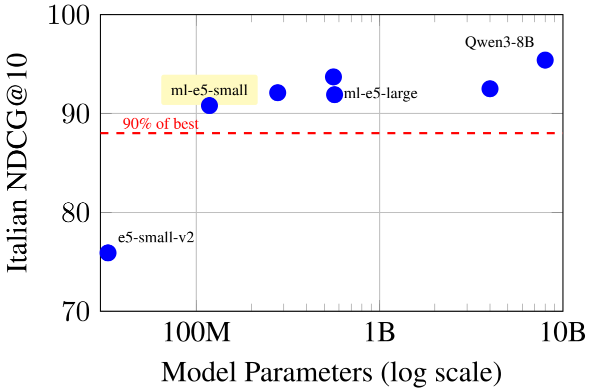

Figure 1: Model size vs. Italian retrieval performance. `ml-e5-small` (118M) achieves 95% of the best model's performance with 68× fewer parameters. Dashed line indicates 90% of the best-performing model

Domain Mismatch. Our training data (mC4 web text) differs substantially from evaluation data (Wikipedia/Belebele encyclopedic content). The model specializes to web-style queries at the cost of encyclopedic retrieval, explaining the 8–15 point NDCG drops.

Hard Negative Circularity. Mining hard negatives with `paraphrase-multilingual-MiniLM-L12-v2` creates a biased training signal. The negatives are "hard" precisely because a similar model cannot distinguish them from positives. Training on these reinforces existing model behavior rather than improving generalization.

4.4 Model Size vs. Performance Trade-off

Figure 1 illustrates the practical implications of our findings.

Efficiency Analysis. For Italian retrieval:

- `ml-e5-small` (118M, 384-dim): 90.8 NDCG@10
- `Qwen3-8B` (8B, 4096-dim): 95.4 NDCG@10

The 4.6-point gap comes at the cost of 68× more parameters and 10.7× larger embedding dimension, translating to significantly higher inference costs, memory requirements, and vector storage overhead.

We note that our interpretation of performance saturation and forgetting is conditioned on the choice of training data, negative mining strategy, and evaluation benchmarks. Alternative fine-tuning approaches (e.g., domain-matched corpora or regularization-based methods) may yield different dynamics.

5 Discussion

Implications for Practitioners. Our findings suggest a simple but impactful change to model selection workflows: **always filter MTEB results by your target language** before making deployment decisions. The aggregated leaderboard score can mislead by up to 13+ points for specific languages.

When to Fine-tune. Language-specific fine-tuning appears counterproductive when:

1. Base model NDCG $>$ 0.90 on target tasks (saturation).
2. Training domain differs from evaluation domain.
3. Hard negatives are mined with the target model family.

Broader Implications. While we focus on Italian, this aggregation issue may affect other languages, particularly those with limited task coverage. Languages with fewer MTEB tasks may show larger discrepancies, as their influence on the aggregate score is diluted.

6 Conclusion

We demonstrate that MTEB's cross-language aggregation masks language-specific performance, causing practitioners to underestimate small multilingual models. For Italian retrieval, `ml-e5-small` (118M parameters) achieves 90.8 NDCG@10 only 2.9 points below 560M parameter models despite showing a 7-point gap on aggregated MTEB scores.

Fine-tuning such saturated models causes catastrophic forgetting (11–49% drops) under our fine-tuning setup, even with extensive hyperparameter tuning, confirming that additional Italian-specific training provides no benefit.

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. Small multilingual models offer excellent cost-performance trade-offs for language-specific applications.

We release the first large-scale Italian retrieval dataset (25.4M triplets) and a domain-specific RAG benchmark to support future research.

Future Work

Cross-Lingual Analysis. While we focus on Italian, the MTEB aggregation artifact likely affects other languages. A systematic study across multiple languages (e.g., German, French, Spanish, Arabic) would reveal whether the discrepancy patterns we observe generalize and identify which languages suffer most from aggregation bias.

Domain-Matched Fine-Tuning. Our catastrophic forgetting results may stem from domain mismatch between mC4 web text and encyclopedic MTEB tasks. Future work should investigate fine-tuning on domain-matched corpora (e.g., Italian Wikipedia, legal documents) to determine whether forgetting persists when training and evaluation domains align.

Alternative Fine-Tuning Strategies. We employed standard contrastive learning with InfoNCE loss. Exploring regularization-based approaches (e.g., elastic weight consolidation, knowledge distillation from base models) might mitigate catastrophic forgetting while preserving domain-specific gains.

Task Diversity Analysis. MTEB Italian contains only two retrieval tasks. As MTEB expands to include more Italian tasks (classification, clustering, semantic similarity), analyzing whether aggregation bias persists across task types would strengthen our conclusions.

Deployment Case Studies. Real-world deployment comparisons between small multilingual models and large-scale alternatives would quantify practical trade-offs beyond NDCG scores, including latency, throughput, storage costs, and end-to-end RAG system performance.

Limitations

Language Scope. Our analysis focuses exclusively on Italian. While we hypothesize similar aggregation effects exist for other languages, we have not empirically verified this.

Task Coverage. MTEB contains only two Italian retrieval tasks (BelebeleRetrieval, WikipediaRetrievalMultilingual). Our conclusions about Italian performance are limited by this task coverage.

Domain Specificity. The observed catastrophic forgetting may be partially explained by domain mismatch between mC4 (web text) and MTEB (encyclopedic). Results might differ with domain-matched training data.

Hard Negative Mining. We used `paraphrase-multilingual-MiniLM-L12-v2` for hard negative mining rather than the target models. Using target models might yield different fine-tuning dynamics.

Hyperparameter Search. While we conducted 8 Optuna trials, a more extensive search might identify configurations that avoid catastrophic forgetting, though the consistency of degradation across all trials suggests this is unlikely.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

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.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-07-18.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Rageval: Scenario specific rag evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262*.

A Extended MTEB Results

Table 4 shows complete results for all evaluated models on Italian MTEB tasks.

B Custom RAG Benchmark Extended Results

Table 5 shows results at multiple k values.

C Hyperparameter Search Details

We conducted 8 Optuna trials exploring:

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

All 8 configurations resulted in performance degradation on both MTEB Italian tasks and our custom RAG benchmark. This consistency across diverse hyperparameters strongly suggests task saturation rather than suboptimal configuration.

Model	Params	Dim	MTEB_{agg}	Belebele	Wiki_{IT}	Avg_{IT}	Gap
Small Models (<200M parameters):							
ml-e5-small	118M	384	77.01	92.41	89.25	90.83	**+13.82**
e5-small-v2	33M	384	41.60	74.08	77.73	75.91	+34.31
bilingual-small	117M	384	78.20	90.00	85.86	87.93	+9.73
Medium Models (200M-600M parameters):							
ml-e5-base	278M	768	79.74	94.18	89.93	92.06	+12.32
bilingual-base	278M	768	81.25	93.24	88.03	90.63	+9.38
ml-e5-large	560M	1024	84.15	95.08	92.34	93.71	+9.56
bilingual-large	559M	1024	84.62	94.33	89.61	91.97	+7.35
bge-m3	568M	1024	84.02	93.67	90.18	91.93	+7.91
snowflake-arctic-l	568M	1024	82.26	93.71	91.75	92.73	+10.47
jina-v3	572M	1024	81.27	93.33	89.18	91.25	+9.98
bge-fr-en	595M	1024	82.68	91.66	90.04	90.85	+8.17
Qwen3-0.6B	595M	1024	77.94	92.83	88.42	90.62	+12.68
Large Models (4B-8B parameters):							
Qwen3-4B	4B	2560	86.19	95.25	89.76	92.50	+6.31
Qwen3-8B	8B	4096	90.39	98.69	92.10	95.39	+5.00
7B Instruct Models:							
SFR-Mistral	7B	4096	80.29	94.48	93.09	93.78	+13.49
e5-mistral-7b	7B	4096	79.61	94.18	92.73	93.46	+13.85
SFR-2-R	7B	4096	77.90	92.60	92.43	92.52	+14.62

Table 4: Extended MTEB Italian retrieval results (NDCG@10). MTEB_{agg} is the official aggregated score across all languages; Avg_{IT} averages BelebeleRetrieval (Italian) and WikipediaRetrievalMultilingual (Italian monolingual). Gap shows the discrepancy (Avg_{IT} - MTEB_{agg}). The highlighted row shows ml-e5-small with a 13.82-point gap, demonstrating how aggregation systematically undervalues small multilingual models.

D Resources

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

Model	k	R@k	NDCG	MRR
Qwen3-8B	1	0.203	0.401	0.401
Qwen3-8B	10	0.466	0.535	0.496
Qwen3-8B	100	0.737	0.542	0.505
ml-e5-small	1	0.181	0.344	0.344
ml-e5-small	10	0.445	0.488	0.445
ml-e5-small	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	1	0.139	0.267	0.267
e5-small-v2	10	0.367	0.408	0.363
e5-small-v2	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	0.361	0.325
BM25	100	0.605	0.402	0.336

Table 5: Custom Italian RAG benchmark results at various k values. Fine-tuned models show consistent degradation across all k.