Linguistic Specialization of Arabic in Text Embedding Models via Multi-Teacher Knowledge Distillation

Ahmad Abdelfattah
Misraj.AI
Syria
ahmad.abdalfattah@misraj.ai

Zeina Aldallal
Misraj.AI
Syria
aldallal@misraj.ai

Sara Chrouf
Misraj.AI
Saudi Arabia
chrouf@misraj.ai

Safwan AlModhayan
Misraj.AI
Saudi Arabia
safwan@misraj.ai

Abstract

Knowledge distillation has proven to be a powerful paradigm for compressing large neural networks into compact yet capable models, embedding models being one such example. In this work, we investigate the use of ensemble-based teacher distillation to specialize a multilingual model for the Arabic language. We demonstrate that aggregating signals from multiple teacher models is a more effective strategy than the conventional single-teacher approach that is predominantly followed in current practice. Our findings further confirm that knowledge distillation serves as a general and reliable mechanism for enhancing the capabilities of embedding models for specific downstream tasks for the target language, with the resulting model's performance being directly shaped by the collective of the employed teachers. Building on these insights, we introduce *kawn-embed-light*, a lightweight Arabic embedding model that achieves state-of-the-art performance among models under 1 billion parameters on the Arabic information retrieval tasks of the Multilingual Massive Text Embedding Benchmark (MMTEB), reaching an average NDCG@10 (Normalized Discounted Cumulative Gain at rank 10) score of **0.7711**. Notably, *kawn-embed-light* remains competitive with significantly larger models and surpasses the majority of closed-source commercial embedding APIs. Finally, our work demonstrates that ensemble distillation provides a computationally efficient and high-performance pathway for model localization, offering a scalable solution to enhance Natural Language Processing (NLP) capabilities in low-resource languages.

1 Introduction

In Natural Language Processing (NLP), sentence embedding refers to a numeric representation of a sentence in the form of a fixed-length vector of real numbers, that encode meaningful semantic information. These embeddings can then be used in various downstream tasks, such as information retrieval, semantic text similarity, classification, clustering, and Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020).

Recently, there has been a growing focus on utilizing knowledge distillation in the context of embedding models (Zhao et al., 2025; Akram et al., 2026; Zhang et al., 2025a, 2024; Kim et al., 2023; Vera et al., 2025). Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) provides an efficient mechanism for training smaller models to mimic the behavior of one or more teacher models (Akram et al., 2026).

Nevertheless, the application of knowledge distillation for acquiring or improving language-specific capabilities in embedding models remains largely unexplored. Moreover, Arabic remains underrepresented in embedding research; dedicated Arabic embedding models remain scarce, and the best-performing Arabic models on Multilingual Massive Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025) are multilingual with parameter counts exceeding 4B (Zhang et al., 2025b; Babakhin et al., 2025) rendering them impractical for resource-constrained and production deployment scenarios due to their prohibitive computational cost.

In this work, we present a novel approach for dis-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 459–467
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_047

tilling Arabic embedding knowledge from multiple large multilingual embedding models (i.e., teachers) into a smaller multilingual embedding model (i.e., student). We demonstrate that multi-teacher distillation (MTD) outperforms the commonly used single-teacher approach.

Our contributions are summarized as follows:

- We leverage Knowledge Distillation to enhance the Arabic language capabilities of embedding models, eliminating the need for expensive from-scratch training.
- We introduce a novel linguistic specialization through MTD approach for embedding models and empirically show that distilling from multiple teachers outperforms relying on the single best teacher alone.
- We introduce *kawn-embed-light*, a 300M-parameter Arabic embedding model that achieves state-of-the-art retrieval performance among sub-1B-parameter models on the Multilingual Massive Text Embedding Benchmark (MMTEB), with an average NDCG@10 (Normalized Discounted Cumulative Gain at rank 10) of 0.7711. This result validates our proposed multi-teacher distillation (MTD) approach for adapting embedding models to new languages.

2 Related Work

2.1 Text Embedding Models

Reimers and Gurevych (2019) proposed an early approach that leverages Siamese networks without contrastive learning. Subsequently, researchers introduced two-stage contrastive learning frameworks built on encoder-only transformers, where the first stage pre-trains on large-scale weakly supervised data and the second stage fine-tunes on smaller, higher-quality supervised datasets (Wang et al., 2022; Li et al., 2023; Sturua et al., 2024; Xiao et al., 2024).

Another line of research explores decoder-only transformers for text embedding. Some works incorporate a pre-training phase on synthetic data (Lee et al., 2025; Zhang et al., 2025b; Babakhin et al., 2025; Zhao et al., 2025), while others skip this phase entirely and apply direct fine-tuning (Hu et al., 2025; Wang et al., 2024a). The latter demonstrate that pre-training provides no significant benefit for decoder-only models, attributing this to their strong autoregressive capabilities.

Despite their differences, both approaches may rely on large scale weakly supervised data and always require high-quality supervised datasets for fine-tuning. For Arabic, such resources are extremely limited, making this a significant challenge.

2.2 Embedding Model Distillation

Knowledge distillation has been applied to embedding models through several complementary strategies. Yang et al. (2024) explore distillation for dense retrieval in a cross-lingual setting, using machine translation to transfer retrieval knowledge across languages.

(Kim et al., 2023) introduce a foundational approach for single-teacher embedding distillation, in which the student is trained to match the teacher's embedding space via an alignment objective, with a projection layer added on top of the student to handle any dimensional mismatch between the two models. Building on a similar single-teacher paradigm, Akram et al. (2026) apply distillation as a pre-training phase using a standard cosine-similarity objective, demonstrating that embedding supervision can be incorporated even before task-specific fine-tuning. They also follow this distillation stage with task-adapter fine-tuning.

A complementary line of work extends distillation to multiple teachers. Zhang et al. (2024) propose concatenating the embedding vectors of several teachers into a single target vector, training the student to reproduce this combined representation through both score-based distillation and embedding alignment, again relying on a projection layer to accommodate the enlarged target dimensionality. Zhang et al. (2025a) follow a similar approach, concatenating two teacher embeddings and optimizing the student with a cosine-similarity objective, demonstrating that ensemble distillation scales effectively to larger models.

On the other hand, Zhao et al. (2025) utilize distillation as a post-training phase, applied after an initial contrastive learning stage. Rather than minimizing cosine distance directly, they distill the normalized distribution of temperature-scaled cosine similarity scores via KL-divergence, transferring richer relational information encoded in the teacher's score distribution.

However, embedding space distillation is not well-suited for multi-teacher distillation because the differing embedding dimensions across multi-

ple teachers make projection-based alignment suboptimal, as no single projection can faithfully map the student to all teacher spaces. Furthermore, there is limited research on the use of distillation techniques to improve the capabilities of the embedding model in a specific language.

3 Dataset

This section describes the dataset used to train our Arabic embedding model. We detail how the data was collected from multiple complementary sources, how hard negatives were mined to construct effective training triplets, and the preprocessing steps applied to ensure data quality. The final dataset comprises approximately 2M training examples drawn from more than 20 distinct sources, spanning general, Islamic, and medical domains.

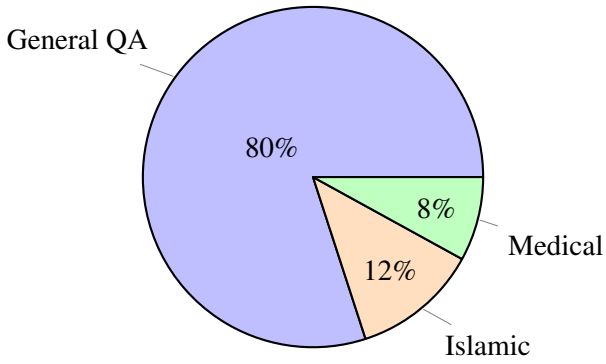

Figure 1: Distribution of data domains across the dataset. General QA accounts for the majority (80%), with Islamic (12%) and Medical (8%) content comprising the remainder.

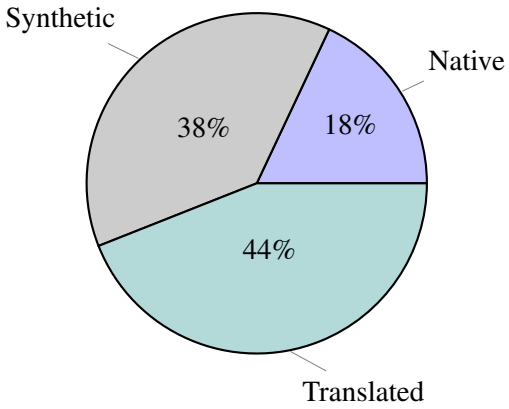

Figure 2: Distribution of data sources across the dataset. Translated data constitutes the largest share (44%), followed by Synthetic (38%) and Native (18%).

3.1 Data Collection

The training data used to train the embedding model is formed as triplets $\{q_i, d_i^+, \{d_{i,k}^-\}_{k=1}^n\}$, where d_i^+ denotes the positive document associated with query q_i, and $\{d_{i,1}^-, \ldots d_{i,n}^-\}$ represents a set of n hard negative (irrelevant) documents of q_i.

However, Arabic remains a low-resource language in this context, with very limited publicly available datasets conforming to this training scheme. To address this limitation, prior work has proposed several practical solutions, including translating existing datasets from high-resource languages into the target low-resource language (Babakhin et al., 2025; Sturua et al., 2024), or generating synthetic datasets by prompting large language models (LLMs) to produce queries for existing documents an approach that has gained considerable adoption in recent years (Babakhin et al., 2025; Hu et al., 2025; Zhang et al., 2025b; Wang et al., 2024a).

Guided by these strategies, our dataset construction follows four complementary directions: (1) leveraging existing Arabic datasets, (2) translating widely adopted information retrieval benchmarks including MS-MARCO (Bajaj et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and SQuAD (Rajpurkar et al., 2016) into Arabic, (3) scraping a large volume of Arabic Islamic community question-answering content from online resources, and (4) generating synthetic data following the approach proposed by (Zhang et al., 2025b). Figure 1 illustrates the domain-based distribution of the compiled dataset. Approximately 80% of the data corresponds to general question-answering, encompassing translated, natively Arabic, and synthetically generated samples. Islamic-domain content accounts for 12% of the dataset, sourced entirely from scraped online resources, while the remaining 8% pertains to the medical domain.

From the perspective of data sourcing, Figure 2 reveals that 44% of the dataset is translated from well-established English information retrieval benchmarks, 38% is synthetically generated using Gemini 2.5 Flash and Gemini 2.0 Flash, following the approach introduced by Zhang et al. (2025b) by generating synthetic queries for documents using a two-stage prompting technique (in our case, using an initial corpus scraped from native Arabic websites covering diverse domains). Concretely, the first stage is a *configuration* stage, where the model is prompted to select a query type (e.g., key-

word, factual, summary), a difficulty level, and a persona likely to be interested in the given document; the second stage is a *query generation* stage, where these configuration choices condition the generation of a realistic query targeting that document. This persona-driven, multi-dimensional conditioning encourages diversity and realism in the resulting query distribution. The remaining 18% consists of natively Arabic content that is either sourced from open datasets or scraped from Islamic QA communities.

3.2 Hard Negatives Mining

To mine hard negatives, we utilize the positive-aware approach proposed by (Moreira et al., 2024); specifically, the `TopK-PercPos` strategy. This strategy retrieves top-k most similar documents for each query, then excludes passages whose similarity score exceeds a given percentage of the query-positive score, in order to eliminate false-negative cases. For the positive-aware mining algorithm, the similarity threshold was set to 95% of the query-positive score.

3.3 Data Pre-processing

Utilizing the embedding that we calculated in the hard negative mining step we filter out pairs whose query-positive similarity falls below 0.35, as well as pairs containing documents below a minimum length threshold, as these are considered noisy or uninformative training examples. After this cleaning process, we retain approximately 2M diverse training examples spanning more than 20 distinct sources.

4 Methodology

In this section, we present our methodology, which is based on knowledge distillation. Formally, let S denote a multilingual sentence embedding model (the *student*), and let $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$ denote a set of K stronger multilingual embedding models (the *teachers*). The method uses knowledge distillation to transfer Arabic language understanding from multiple teacher models into a single student model by minimizing the KL divergence between the soft probability distributions of the ensemble teachers and the student model.

Inspired by (Zhao et al., 2025), we distill normalized temperature-scaled cosine similarity scores into the student model. Our method is built upon two key design choices: (1) we employ a single unified training phase, eliminating the need for a separate post-contrastive fine-tuning stage in which the model would first be trained with a standard contrastive loss and subsequently fine-tuned; and (2) we leverage an ensemble of multiple teacher models rather than relying on a single best-performing teacher.

Formally, Training data consists of triplets $\{q_i, d_i^+, \{d_{i,j}^-\}_{j=1}^n\}$ (Section 3). For each sample, a model M scores every candidate document $d_i \in \{d^+\} \cup \{d_1^-, \ldots, d_n^-\}$ as a normalized cosine similarity: $z_i^M = \phi(M(q_i), M(d_i))$, where ϕ denotes the cosine similarity function. These scores form a logit vector $\mathbf{z}^M = (z_1^M, \ldots, z_N^M) \in \mathbb{R}^N$, with $N = 1 + n$, over which a temperature-τ softmax is applied to produce candidate probabilities, as defined in Equations 1 and 2 below.

The individual teacher probabilities $P_{t,k}(i)$ and the student probabilities $P_S(i)$ are obtained by applying a temperature-scaled softmax transformation over their respective similarity logits, where $\tau > 0$ is a temperature hyperparameter controlling the smoothness of the resulting distributions. For the student model with similarity logits $z_{s,i}$:

$$P_S(i) = \frac{\exp(z_{s,i}/\tau)}{\sum_{j=1}^{N} \exp(z_{s,j}/\tau)}, \tag{1}$$

and for the k-th teacher with similarity logits $z_{t,k,i}$:

$$P_{t,k}(i) = \frac{\exp(z_{t,k,i}/\tau)}{\sum_{j=1}^{N} \exp(z_{t,k,j}/\tau)}. \tag{2}$$

Higher values of τ produce softer distributions that emphasize relative ranking across all candidates, while lower values produce sharper distributions that concentrate probability mass on the highest-scoring candidate.

The learning objective is the Kullback–Leibler (KL) divergence between the aggregated teacher distribution $P_{\mathcal{T}}$ and the student distribution P_S. We adopt the direction $D_{KL}(P_{\mathcal{T}} \parallel P_S)$, treating the teacher ensemble as the reference distribution and penalizing the student for deviating from it:

$$\begin{aligned} \mathcal{L}_{KL} &= D_{KL}(P_{\mathcal{T}} \parallel P_S) \\ &= \sum_{i=1}^{N} P_{\mathcal{T}}(i) \log\left(\frac{P_{\mathcal{T}}(i)}{P_S(i)}\right). \end{aligned} \tag{3}$$

The aggregated target distribution $P_{\mathcal{T}}$ is constructed as a convex combination of the K individual teacher distributions, with each teacher's

contribution is modulated by a scalar weight w_k:

$$P_{\mathcal{T}}(i) = \sum_{k=1}^{K} w_k P_{t,k}(i) \quad \text{where} \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0. \tag{4}$$

This formulation guarantees that $P_{\mathcal{T}}$ constitutes a valid probability distribution. The weights $\{w_k\}$ thus provide fine-grained control over each teacher's relative contribution to the distillation signal.

5 Experiments and Results

5.1 Setup

The student model is initialized from *EmbeddingGemma-300M* (Vera et al., 2025), selected for its optimal trade-off between computational efficiency and performance on Arabic information retrieval tasks on MMTEB. Knowledge distillation is performed using two high-capacity multilingual teacher models: *Qwen3-Embedding-8B* (Zhang et al., 2025b) and *Llama-Embed-Nemotron-8B* (Babakhin et al., 2025), both of which rank among the strongest publicly available models for Arabic information retrieval tasks on the benchmark. Training proceeds for approximately 15,000 optimization steps over a corpus of 2 million high-quality, filtered instances with 4 hard negatives per query-document pair, with context length 1024 token. All experiments run on 2 × NVIDIA H100 (80GB) GPUs with a global batch size of 128. We use BFloat16 (bf16) mixed-precision to ensure numerical stability while reducing memory overhead, and optimize with a learning rate of 1×10^{-5} and a linear warmup schedule over the first 100 steps. For softmax temperature scaling we use $\tau = 1$. Following the standard protocol established for *EmbeddingGemma-300M*, we apply the prescribed prompt templates to both queries and documents during training. For knowledge distillation, we use a uniform ensemble weighting strategy, assigning equal weight to each teacher signal ($w_k = 0.5$ per teacher), and treating both teachers as equally reliable.

5.2 Single-Teacher vs. Multi-Teacher Distillation

To determine the optimal distillation strategy, we compare three configurations: distilling from *Qwen3-Embedding-8B* alone, from *LLaMA-Embed-Nemotron-8B* alone, and from the ensemble of both teachers (proposed approach). As shown in Table 1, ensemble distillation outperforms both single-teacher baselines, including the stronger *LLaMA-Embed-Nemotron-8B* teacher, demonstrating that combining multiple teachers yields richer training signal than any individual teacher alone. Furthermore, the single-teacher results exhibit a consistent ordering: distilling from *LLaMA-Embed-Nemotron-8B* yields superior performance compared to distilling from *Qwen3-Embedding-8B*, which aligns with the relative performance gap between the two teachers observed in Table 2. All results are reported on the MMTEB benchmark using the evaluation tasks described in Section 6.1.

Configuration	Avg
Teacher Models (Upper Bound)	
LLaMA-Embed-Nemotron-8B	0.7980
Qwen3-Embedding-8B	0.7883
Student Model (Lower Bound)	
EmbeddingGemma-300M	0.7259
Single-Teacher Distillation	
w/ LLaMA-Embed-Nemotron-8B	0.7626
w/ Qwen3-Embedding-8B	0.7565
Multi-Teacher Distillation	
kawn-embed-light (proposed)	**0.7711**

Table 1: Comparison of distillation configurations. Our proposed multi-teacher distillation (MTD) outperforms all single-teacher baselines despite using a significantly smaller student model (*kawn-embed-light*, 300M parameters).

6 Evaluation

6.1 Evaluation Settings

For evaluation we utilize Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025), encompasses a comprehensive collection of evaluation tasks spanning over 250 languages and ten task types: Bitext Mining, Classification, Clustering, Instruction Retrieval, Multilabel Classification, Pair Classification, Reranking, Retrieval, Semantic Textual Similarity (STS), and Summarization.

As this work focuses on distilling information retrieval capabilities for an Arabic language model,

Model	Number Of Parameters(M)	Avg	MRHN	MLQA
Commercial APIs				
google/gemini-embedding-001	-	**0.7934**	**0.7877**	**0.7990**
Cohere/Cohere-embed-multilingual-v3.0	-	0.7445	0.7679	0.7210
voyageai/voyage-3	-	0.7201	0.7012	0.7390
openai/text-embedding-3-large	-	0.6812	0.6912	0.6712
Large Models (More than 4B Params)				
nvidia/llama-embed-nemotron-8b	8000	**0.7980**	**0.8100**	**0.7859**
Qwen/Qwen3-Embedding-8B	8000	0.7883	0.7921	0.7845
Small Models (Less than 1B Params)				
kawn-embed-light	300	**0.7711**	**0.7896**	**0.7527**
intfloat/multilingual-e5-large	596	0.7497	0.7658	0.7336
BAAI/bge-m3	568	0.7458	0.7890	0.7026
jinaai/jina-embeddings-v5-text-small	596	0.7404	0.7480	0.7328
google/embeddinggemma-300m	300	0.7259	0.7278	0.7239
Qwen/Qwen3-Embedding-0.6B	595	0.6829	0.7050	0.6608

Table 2: Arabic MMTEB Retrieval Tasks Evaluation Results. MRHN: MIRACL Retrieval Hard Negatives, MLQA: MultiLingual Question Answering

we leverage two prominent retrieval tasks from the benchmark, selected based on their prevalence among submissions by compared model authors. MIRACL Retrieval Hard Negative which is modified more compatitive version of MIRACL (Zhang et al., 2023), and MLQA. For both tasks, we adopt **NDCG@10** (Normalized Discounted Cumulative Gain at rank 10) as the evaluation metric, consistent with MMTEB's default scoring protocol for information retrieval tasks.

6.2 Main Results

We compare our model with various open-source text embedding models and commercial API services. The open-source models include two of the best Multilingual embedding models larger than 4B parameters: Llama-Embed-Nemotron-8B (Babakhin et al., 2025) and Qwen3-Embedding-8B (Zhang et al., 2025b), which also serve as our teacher models in the knowledge distillation process. For small models less than 1B parameters, we include Multilingual-E5-Large (Wang et al., 2024b), BGE-M3 (Chen et al., 2024), Jina-Embeddings-v5-Text-Small (Akram et al., 2026), EmbeddingGemma-300M (Vera et al., 2025), and Qwen3-Embedding-0.6B (Zhang et al., 2025b). The commercial APIs evaluated are Gemini-Embedding-001 from Google (Lee et al., 2025), Voyage3 from Voyage AI, Cohere-Embed-Multilingual-v3.0 from Cohere, and Text-Embedding-3-Large from OpenAI. For all models, we use results submitted by their authors to the MMTEB results repository[1].

Table 2 presents the evaluation results on Arabic MMTEB retrieval tasks, covering two tasks: MIRACL Retrieval Hard Negatives (MRHN) and MultiLingual Question Answering (MLQA). We compare *kawn-embed-light* against commercial APIs, large-scale open models (>4B parameters), and small open models (<1B parameters).

Among the commercial APIs, Gemini-Embedding-001 achieves the highest average score of 0.7934, setting a strong competitive ceiling. In the large model category, Llama-Embed-Nemotron-8B and Qwen3-Embedding-8B which serve as the teacher models in our ensemble distillation framework achieve average scores of 0.7980 and 0.7883, respectively, yielding a mean teacher performance of 0.7932.

Our proposed model, *kawn-embed-light*, is a compact 300M parameter model trained via ensemble knowledge distillation from both teacher models, using EmbeddingGemma-300M as the student backbone. Despite its significantly reduced size more than 26× smaller than each teacher *kawn-embed-light* achieves an average score of 0.7711, corresponding to 97.2% of the mean teacher performance. Compared to the base student model before distillation, EmbeddingGemma-300M, which scores 0.7259, *kawn-embed-light* yields an absolute improvement of +4.52%, demonstrating the substantial gain introduced by the ensemble dis-

[1]See the MMTEB results repository at https://github.com/embeddings-benchmark/results.

Model	Avg. Retr.
kawn-embed-light	**0.7711**
Merge-Distill (Qwen × LLaMA)	0.7634

Table 3: Model merging vs. proposed approach on average retrieval.

Model	Avg. STS
kawn-embed-light	**0.7327**
embeddinggemma-300m	0.7204

Table 4: Semantic Textual Similarity (STS) benchmark results on the Arabic splits of STS-17 and STS-22. Spearman rank correlation is computed between model-predicted and ground-truth similarity scores. *embeddinggemma-300m* denotes the pre-fine-tuning baseline.

tillation process. Notably, *kawn-embed-light* outperforms all other small models (less than 1B params) by a substantial margin, surpassing the second-best small model, Multilingual-E5-Large (596M), by +2.14% in average score, while also exceeding several closed-source baselines including Cohere-Embed-Multilingual-v3.0 from Cohere, Voyage3 from Voyage AI, and Text-Embedding-3-Large from OpenAI.

7 Analysis and Discussion

In this section, we analyze two key aspects of our proposed approach. First, we investigate whether model merging can serve as a viable alternative to ensemble distillation. Second, we assess whether the distillation process causes catastrophic forgetting by evaluating the model on semantic textual similarity (STS) as a proxy for general language understanding.

7.1 Model Merging

Inspired by recent works applying model merging as a post-training technique (Zhang et al., 2025b; Lee et al., 2025; Vera et al., 2025), we investigate whether merging separately distilled models can yield competitive or superior performance compared to our proposed ensemble distillation approach. To this end, we apply Spherical Linear Interpolation (SLERP) to merge the two individually distilled teacher-based models described previously. As reported in Table 3, the ensemble distillation approach still yields substantial improvements over the merged model, demonstrating that joint optimization from multiple teachers during training captures richer complementary signals than post-hoc weight interpolation.

7.2 Catastrophic Forgetting and General Knowledge Improvement

To assess whether the distillation process degrades the student model's pre-existing knowledge given that training was conducted exclusively on information retrieval data we evaluate the resulting model on a semantically distinct task type: Semantic Textual Similarity (STS). Specifically, we benchmark on two widely adopted Arabic STS datasets, STS17 and STS22, both of which are available within the MMTEB benchmark suite. As shown in Table 4, despite the complete absence of STS data during training, the distilled model achieves an improvement of +1.2% over the original EmbeddingGemma-300M baseline. This result provides two key insights: first, it confirms that the distillation procedure does not induce catastrophic forgetting of previously acquired capabilities; second, it suggests that the process yields a broader improvement in the model's general understanding of Arabic, extending beyond the information retrieval domain.

8 Conclusion

In this work, we presented *kawn-embed-light*, a compact 300M-parameter Arabic embedding model trained for information retrieval via multi-teacher knowledge distillation from *Qwen3-Embedding-8B* and *Llama-Embed-Nemotron-8B*. Despite its small size, *kawn-embed-light* outperforms all evaluated small models and several commercial APIs on Arabic MMTEB retrieval benchmarks. Our experiments demonstrate that ensemble distillation consistently surpasses single-teacher distillation, and STS evaluation confirms that the distillation process enhances general Arabic language understanding without causing catastrophic forgetting o overcome the scarcity of Arabic IR training data, we constructed a diverse corpus by combining LLM-based translation, synthetic data generation, and web scraping. These results highlight the potential of knowledge distillation as an effective strategy for building efficient, high-quality Arabic embedding models..

References

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, et al. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. 2023. Embeddistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

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.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

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. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *European Conference on Information Retrieval*, pages 50–65. Springer.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Dun Zhang, Ziyang Zeng, Yudong Zhou, and Shuyang Lu. 2025a. Jasper-token-compression-600m technical report. *arXiv preprint arXiv:2511.14405*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

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

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, et al. 2025. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *arXiv preprint arXiv:2506.20923*.