MORSE: MIXTURE OF REPRESENTATIONS FROM SPARSIFIED EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on quantization (*e.g.,* TurboQuant) and dimensionality reduction (*e.g.,* Matryoshka Representation Learning; MRL) demonstrates that text embeddings contain substantial representational redundancy. However, enforcing extreme sparsity (retaining only a small number of active dimensions) can reduce embedding generality and favor task-specific performance, as observed in work on Contrastive Sparse Representations (CSR) and CSRv2. We show that such task-specific, low-*active* dimension representations can be learned by training a lightweight adapter (namely, *Sparsified Expert*) over any base model using a novel weighted-top-k objective. Our approach departs from prior methods that either (1) concentrate information in predetermined leading dimensions (*i.e.,* prefix-based), as in MRL and Matryoshka Adapters, or (2) project embeddings into a higher-dimensional space to identify sparse representational axes, as in CSR and CSRv2. We then introduce *Mixture of Representations from Sparsified Experts* (`MORSE`), which combines representations from multiple `SE`s to recover task-agnostic performance while maintaining sparsity in representations. `MORSE` occupies the previously underexplored regime between moderately sparse, task-agnostic representations and extremely sparse, task-specific ones, revealing a Pareto frontier between representational sparsity and generality. Across three base models, our `SE` outperforms other methods at ultra-low dimensions and `MORSE` improves over others on cross-lingual generality.

1 INTRODUCTION

Modern AI systems increasingly depend on retrieval throughout inference. In particular, agentic systems repeatedly access external knowledge, memories, skills, and tools through search indexes that grow over time. While high-dimensional representations have demonstrated good performance, the question of whether the representation is compact to capture all salient information was unclear. Further, the larger representations incurred increased storage and computational latency costs. To address these, existing works have considered reducing the representation dimension (Kusupati et al., 2022; Yoon et al., 2024) or compressing information in a representation via rotations, and quantization (Ge et al., 2014; Zandieh et al., 2025). Interestingly, approaches like Matryoshka Representation Learning (`MRL`) (Kusupati et al., 2022) have demonstrated performant generalized representations are possible with lower dimensions when learned during pretraining. Most embedding models today are trained using the `MRL` paradigm and automatically consider the notion of prefix-based sparsity in the representation space, where the `first-k` dimensions encapsulate most of the information from a d-dimensional base embedding ($d > k$) while retaining reasonable downstream performance.

Recent works have demonstrated we can compress further using post-hoc adapter-based approaches that have slight downstream performance impact (Yoon et al., 2024; Takeshita et al., 2026), but generalized representation quality starts to deteriorate drastically at ultra-low dimensions (see Fig. 1). To recover from this quality degradation, Wen et al. (2025) and Guo et al. (2026) have introduced approaches that learn ultra-sparse representations which outperform `MRL` on specific tasks and datasets, highlighting the trade-off between compression and generality.

This work contributes two methods to this existing landscape. First, as shown in Figure 1 (left), we introduce a lightweight post-hoc adapter (namely, Sparsified Expert) training paradigm that uses a novel weighted-top-k loss function to obtain elastic sparse representations that are more performant than those from existing task-specific experts trained with Matryoshka Adapters (`M-Adaptor`),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

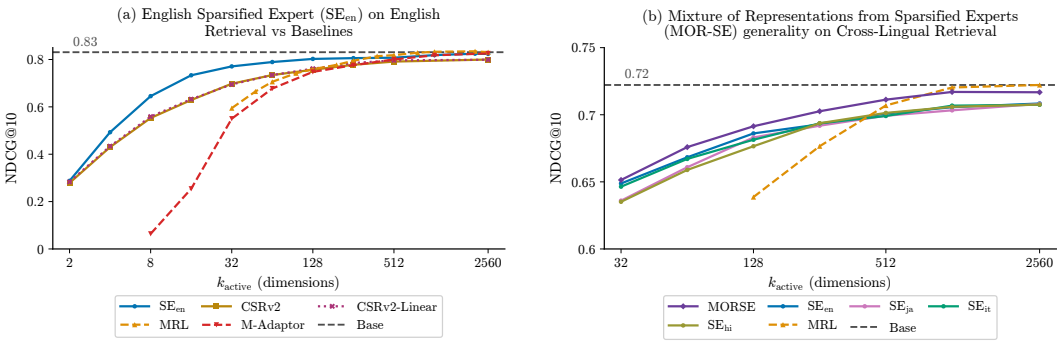


Figure 1: **Specialization and cross-lingual generality with Qwen3-4B.** (a) The English Language Sparsified Expert (SE$_{en}$) achieves the strongest English retrieval performance at compact dimensions from $k_{\text{active}} \in \{2, 4, ..., 256\}$ against the baselines - CSRv2, CSRv2-linear, Matryoshka Representation Learning (MRL), Matryoshka Adapter (M-Adaptor). (b) MORSE generalizes across tasks/domains, *i.e.,* languages - English (en), Japanese (ja), Italian (it), and Hindi (hi), leading individual SE$_{lang}$ and MRL on average cross-lingual retrieval tasks across most budgets and generally approaches the full-dimensional Base as $k_{\text{active}} \in \{32, 64, ..., 512\}$.

CSRv2 (& its variants) and challenge the notion of prefix-based sparsity introduced in MRL. Second, as shown in Figure 1 (right), we introduce a Mixture of Representations from Sparsified Experts (MORSE) that consolidate representations from independently trained SEs into one compact sparse representation that strikes a better trade-off between generality and performance, outperforming MRL at low-dimensions.

2 Related Work

Relevant prior works span three adjacent areas: elastic and compressed embeddings, sparse representations, and expert aggregation.

Elastic and compressed embeddings. MRL trains multiple prefix sizes jointly, concentrating information toward the beginning of the embedding (Kusupati et al., 2022). Each trained prefix can then be used at inference, but this structure must be introduced while training or fine-tuning the encoder. Embeddings trained without MRL can remain competitive under moderate prefix truncation, while MRL becomes more useful under heavy truncation (Takeshita et al., 2026). Quantization reduces the bits used to encode a vector rather than its coordinate count. Optimized product quantization learns a rotation before encoding vector subspaces with codebook identifiers (Ge et al., 2014), while TurboQuant uses randomized rotations and scalar quantizers without encoder retraining (Zandieh et al., 2025). These methods are complementary to the sparse experts, which learn a mapping and store signed values with their coordinate indices.

Sparse representations. Sparse retrieval may use vocabulary-indexed coordinates or a learned latent space. SPLADE learns vocabulary-indexed expansion weights compatible with inverted indexes (Formal et al., 2021). CSR maps frozen dense embeddings into a nonnegative top-k latent space, using $h = 4d$ by default, and trains it with reconstruction and contrastive objectives (Wen et al., 2025). CSRv2 adds target-sparsity annealing and natural task supervision. CSRv2-linear trains only the sparse head, while full CSRv2 also adapts the backbone through LoRA (Guo et al., 2026). The resulting models are trained separately for benchmark task families at ultra-low sparsity. CSR and CSRv2 are the closest representation-level comparisons. Our adapter instead keeps the backbone frozen, uses a signed latent space of dimension d, and jointly optimizes one two-layer sparse autoencoder (SAE) (Makhzani & Frey, 2014) over a top-k ladder. The main experts train only on mMARCO and transfer to MMTEB datasets not used for adapter training, while a separate diagnostic examines the benchmark-specialized retrieval setting of CSRv2 (Appendix E).

Expert aggregation. Embedding expert systems may route inputs among specialized modules or combine their outputs. RouterRetriever selects a domain-specific LoRA expert for each query (Lee et al., 2025), while Mixture of Task Experts (MoTE) routes each task to specialized parameters inside a multitask embedding model (Romero et al., 2025). MORSE differs by mapping final

representations from independently trained experts into one elastic sparse representation shared by queries and a single document index. Appendix B gives additional context.

3 METHODOLOGY

As noted in Section 1, existing approaches enable either sparse representations (Guo et al., 2026) which result in improved downstream performance at low active dimensions or generalized reduced dimension representations (Kusupati et al., 2022) but not both. We therefore question two common paradigms of restricting compact representations to prefixes learned with the base model (*i.e.,* MRL), and increasing representation dimension to induce sparsity (*i.e.,* CSRv2) which are based on information content of dense representations. Instead, we ask whether that information can be reorganized after a backbone has been trained. To this end, we present a two-staged training framework that bridges the gap between above approaches by enabling best of the both worlds *i.e.,* a generalized representation which is on par or better than other low dimensional or sparse representations at both low and high dimensions across tasks.

The proposed framework trains light-weight adapters over a frozen base model (*i.e.,* backbone) in two stages. In the first stage namely, Sparsified Expert (SE) training, we transform the base model's representations into elastic sparse representations (henceforth, ESR) to enhance their performance at low *a*ctive[1] dimensions as existing approaches result in significant drop in performance at $k < 128$ (see Fig. 1). This stage focuses on training individual SEs. In the second stage namely, Mixture of Representations (MoR), we combine the individually learned sparse representations from each SE to obtain a generalized ESR, which we call MORSE. This work focuses on retrieval setting and trains experts on different languages. However, it is crucial to emphasize that the method is generalizable to other domains (*e.g.,* images). We provide details on the two stages in the following sections.

3.1 SPARSIFIED EXPERT TRAINING

Architecture. Wen et al. show that dense representations can be adapted to ultra-sparse representations while retaining reasonable downstream performance. Inspired by this, we train a SE which is a light-weight encoder that takes in normalized version ($z^{(0)}(x) = \frac{z(x)}{\|z(x)\|_2}$) of representation $z(x) \in \mathbb{R}^d$ from a frozen base model and transforms it into a d-dimensional[2] elastic sparse representation (see Fig. 2). To facilitate training, we also use a decoder tied to the encoder and defined as below.

$$f_\theta^{(e)}(z) = W_2\,\mathrm{GELU}(W_1 z), \text{ and } f_\theta^{(d)}(z) = W_1^T\,\mathrm{GELU}(W_2^T z).$$

Here $W_1, W_2 \in \mathbb{R}^{d\times d}$. Note that we do not use the decoder at inference time.

$$\begin{aligned}
z^{(1)}(x) &= f_\theta^{(e)}\left(z^{(0)}(x) - b_{\text{center}}\right) + b_{\text{latent}},\\
\widetilde{z}_{k,i}^{(1)}(x) &= \begin{cases} z_i^{(1)}(x), & i \in \mathcal{I}_k(x),\\ 0, & i \notin \mathcal{I}_k(x),\end{cases},\ \text{with } \mathcal{I}_k(x) = \underset{i\in\{1,\dots,d\}}{\text{argtopk}} \left|z_i^{(1)}(x)\right|,\\
z_k^{(1)}(x) &= \frac{\widetilde{z}_k^{(1)}(x)}{\|\widetilde{z}_k^{(1)}(x)\|_2},\ z_k^{(2)}(x) = f_\theta^{(d)}\left(\widetilde{z}_k^{(1)}(x)\right) + b_{\text{center}}.
\end{aligned} \tag{1}$$

We denote output from the encoder as $z^{(1)}(x)$ and their derived representation with top-k *active* dimensions as $\widetilde{z}_k^{(1)}, \forall k < d$. The decoder takes in the unnormalized representation $\widetilde{z}_k^{(1)}$ to obtain $z^{(2)}(x)$ which is used for computing reconstruction loss (eq. (3)). Here, retrieval encodes a query and candidate documents as normalized embeddings, ranks documents by their cosine similarity (the dot product after normalization), and aims to place relevant documents first. We train each Expert using triplets $\mathcal{B} = \{(q_i, d_i^+, d_i^-)\}_{i=1}^B$ containing a retrieval query, a positive, and a BM25 (Robertson & Zaragoza, 2009) derived negative document using eq. (2).

Optimization. We introduce a novel weighted-top-k objective which is a reconstruction loss regularized with contrastive, and distillation penalizations to train each expert for obtaining ESR (see

[1] Active dimension refers to the number of non-zero values in a sparse representation.

[2] d denotes the full representation dimension of a base model

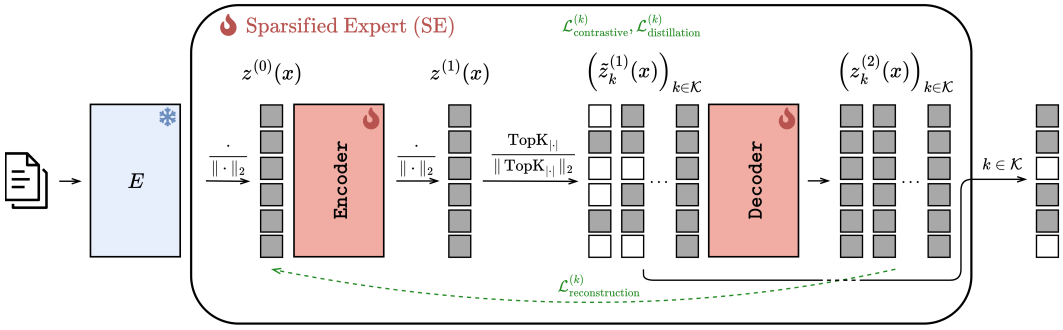


Figure 2: Sparse expert training and inference. A frozen backbone produces dense query and document representations. The trainable encoder maps them to a signed latent representation, top-k yields the elastic sparse representations, and the tied decoder reconstructs the backbone representation. At inference, only the encoder and top-k paths are retained.

eq. (2)).

$$\mathcal{L}_{\text{SE}} = \sum_{k\in\mathcal{K}} w_k \left(\mathcal{L}^{(k)}_{\text{reconstruction}} + \lambda_{\text{contrastive}} \mathcal{L}^{(k)}_{\text{contrastive}} + \lambda_{\text{distillation}} \mathcal{L}^{(k)}_{\text{distillation}} \right) \tag{2}$$

Here **reconstruction loss** aims at preserving the geometry and semantics of the base model's representation into that of learned elastic sparse representation.

$$\mathcal{L}^{(k)}_{\text{reconstruction}} = \frac{1}{3B} \sum_{i=1}^{B} \sum_{x\in\{q_i, d_i^+, d_i^-\}} \left\| z^{(2)}_k(x) - z^{(0)}(x) \right\|_2^2 \tag{3}$$

For a batch of B queries, let $(d_1, \ldots, d_{2B}) = (d_1^+, \ldots, d_B^+, d_1^-, \ldots, d_B^-)$ be the candidate documents. We define the temperature-scaled similarity between a query and a document as

$$s_k(q, d) = \frac{z^{(1)}_k(q)^\top z^{(1)}_k(d)}{\tau}, \text{ where } \tau > 0 \text{ is the temperature}$$

The corresponding query-to-document similarity distribution is $p^{(k)}_{ij} = \text{softmax}_j(s_k(q_i, d_j))$. As multiple candidates may correspond to the same source document. We therefore define the positive set for query q_i as $\mathcal{P}_i = \left\{ j : \text{id}(d_j) = \text{id}(d_i^+) \right\}$. The multi-positive contrastive loss is

$$\mathcal{L}^{(k)}_{\text{contrastive}} = -\frac{1}{B} \sum_{i=1}^{B} \log \left(\sum_{j\in\mathcal{P}_i} p^{(k)}_{ij} \right). \tag{4}$$

Thus, documents sharing the same source as d_i^+ are treated as positives rather than false negatives. When $\mathcal{P}_i = \{i\}$, this objective reduces to the standard InfoNCE loss.

Distillation loss (eq. (5)) aims at ensuring the query-document similarity distribution by using ESR from an expert (student) does not diverge from using that from base model (teacher).

$$\mathcal{L}^{(k)}_{\text{distillation}} = \frac{\tau^2}{B} \sum_{i=1}^{B} D_{\text{KL}} \left(p_i^{(0)} \parallel p_i^{(k)} \right) \tag{5}$$

Our `Expert` training also enables envisioning existing non-MRL or low dimensional embedding models as low dimensional which we show in **??**.

Training Details. We train language-specific `Expert`using their respective samples from `mMARCO` dataset. We use hard-negatives and ensure each minibatch contains both positive and negative documents for a query. Duplicate documents are considered as positives. We use `Muon` (Jordan, 2024) optimizer for matrix parameters and `Adam` (Kingma & Ba, 2015) for vector parameters. Appendices B and C report parameter initialization and count details, and further implementation and weight schedule details, respectively.

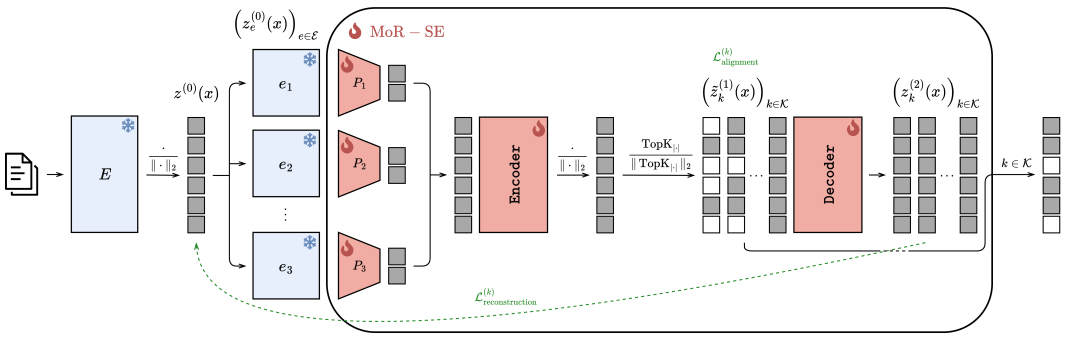


Figure 3: MoR: Frozen expert representations are independently projected, concatenated, and encoded into one signed latent vector. Magnitude top-k produces the stored aggregate, while a training-only decoder supplies reconstruction supervision.

3.2 Mixture of Representations

While sparsified expert is designed to improve performance of sparse representations at ultra low *a*ctive dimensions, they may not generalize across full span of dimensions and across tasks *i.e.,* language in this work (see Fig. 6). To encode generalizability in ESR obtained after Expert training, we propose MoR paradigm which trains a mixture of ESR from each language expert and creates an ensemble ESR (namely, MORSE).

Architecture. Let $z^{(0)}_{e_j}(x) \in \mathbb{R}^d$ denote the normalized full-dimensional representation from expert e_j, for $j \in \{1, \ldots, n\}$. MoR first projects the sparse representation from each expert into a $p = \frac{d}{n}$ dimensional representation using a projection matrix $P_j \in \mathbb{R}^{p \times d}$. The projected representations are concatenated to obtain a d dimensional representation.

$$z^{(1)}(x) = W_e \, \text{concat}_{j=1}^{n} \left(P_j z^{(0)}_{e_j}(x) + b_j \right) + b_e$$

Here $W_e \in \mathbb{R}^{d \times 4p}$. The ensemble uses the same absolute top-k operation and normalization as each expert, with a train time-only decoder:

$$z^{(2)}_k(x) = W_d \tilde{z}^{(1)}_k(x) + b_d$$

Here $W^d \in \mathbb{R}^{d \times d}$. The final ensemble therefore has the same dimension and storage format as one expert.

Optimization. We train MoR using a combination of reconstruction and alignment objectives with a top-k schedule similar to Expert training as defined in eq. (6).

$$\mathcal{L}_{\text{MoR}} = \sum_{k \in \mathcal{K}_{\text{MoR}}} w_k \left(\mathcal{L}^{(k)}_{\text{reconstruction}} + \lambda_{\text{alignment}} \mathcal{L}^{(k)}_{\text{alignment}} \right). \tag{6}$$

Reconstruction loss is defined same as in eq. (3) and serves the same purpose. For each data triple i, we define the frozen expert, ensemble margins, and the alignment loss below:

$$\Delta_{e_j,i} = z^{(0)}_{e_j}(q_i)^T \left(z^{(0)}_{e_j}(d^+_i) - z^{(0)}_{e_j}(d^-_i) \right), \text{ and } \Delta_{\text{MoR}_k,i} = z^{(1)}_k(q_i)^T \left(z^{(1)}_k(d^+_i) - z^{(1)}_k(d^-_i) \right).$$

The alignment penalty is:

$$\mathcal{L}^{(k)}_{\text{alignment}} = \frac{1}{B} \sum_{i=1}^{B} \max \left(0, \max_j \Delta_{e_j,i} - \Delta_{\text{MoR}_k,i} \right)$$

Here, $\Delta_{e_j,i}$ measures how well expert e_j separates the positive document from the negative one for query q_i, and $\Delta_{\text{MoR }k,i}$ is the same for MORSE. Thus at each *active* dimension k, the **alignment loss** pushes MORSE to match its strongest constituent expert on every example, so ensemble never washes out the sharpest expert's signal.

Training Details. We use mMARCO dataset to train MoR on $n = 4$ frozen language experts one each for English, Hindi, Japanese, and Italian. The full d-dimensional ESR are used throughout training and evaluation. The concatenation is done for same input sample across different languages. Further implementation, weighted-top-k schedule, parameter initialization and counts details are provided in Appendices B and C.

4 EXPERIMENTAL SETUP

Given our Elastic Sparse Representations (ESRs) can be trained over any base embedding model, regardless of architecture or whether the training process considered MRL, we consider both decoder-only (Qwen3-Embedding ($0.6B$ and $4B$) (Zhang et al., 2025)) and encoder-only (BGE-M3 (Chen et al., 2024)) multilingual base models in our experiments (see Table A.2 for parameter and embedding dimension details). Specifically, each language expert is trained on respective language's data subset of mMARCO (Bonifacio et al., 2021), a machine-translated multilingual extension of MS MARCO (Bajaj et al., 2018). The English (En), Japanese (Ja), Italian (It), and Hindi (Hi) subsets share same query and documents translated into their respective language.

Baselines. We compare ESR and MORSE to the following setups: (1) full-dimension representation from backbone model, (2) first k dimensions from Qwen3-Embedding (MRL), (3) first k dimensions from Matryoshka Adapters (Yoon et al., 2024) which trains a light-weight adapter over a text embedding model to obtain MRL-like representations (M-Adapter), (4) top-k sparse representation from CSRv2 (Guo et al., 2026) and its LoRA-disabled version (CSRv2-linear), and finally (5) a simple mean of representations obtained from all the experts (Unweighted MORSE).

Data and Evaluation Metrics. We use 5 retrieval datasets per language from MMTEB (Enevoldsen et al., 2025) to evaluate the SEs, MORSE, and the baselines to ensure there is no train-test overlap. We report $NDCG@10$ here and $Recall@10$ in the appendix (Tables F.3 to F.5).

Additionally, to analyze MORSE, we compute a **retention score** to measure the amount of full-dimensional backbone's quality that is preserved after reducing the representation to k dimensions as below:

$$\text{retention}_{M,k,\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{r \in \mathcal{D}} \frac{s_{m,k,r}}{s_{\text{backbone},k_{\text{backbone}},r}}.$$

Here, $\mathcal{D}$ denotes dataset group, $s_{m,k,r}$ refers to the retrieval score of model M at *active* dimension k on dataset r, and k_{backbone} is the full output dimension of the corresponding backbone. The evaluation dataset groups are defined in Appendix A.

5 RESULTS

In this section, we first show that our weighted-top-k loss can create Sparsified Experts (SEs) that are agnostic to the base model (§5.1). Moreover, these SEs show strong task-specific performance in ultra-sparse compression regimes (§5.2). On task-specific metrics, MORSE outperforms SEs beyond a certain number of active-dimensions (k_{active}), and beats other baseline representations (§5.3). On cross-lingual settings, MORSE shows better generalization than SEs, even at low-dimensions (§5.4). Finally, we study the generality of of our SEs on all held-out MMTEB tasks (§5.5).

5.1 MODEL-AGNOSTIC SES

To show that our Sparsified Experts introduces base-model-agnostic ESR, we compare it against (1) MRL when models are natively trained with MRL, (2) First-k, which mimics k_{MRL}, for non-MRL models, such as bge-m3, and (3) M-Adaptor (trained on the same training data of mMARCO). We ignore comparing against CSRv2 on model-agnostic training due to lack of reproducibility, but compare with it in the upcoming sections.

In Fig. 4(a), we see our expert SE_{en} outperforms both M-Adaptor and First-k on every $k_{active} \leq 512$. As shown in Fig. 4(b), we also find that bge-m3 SE_{en} retention curve follows a similar curve as Qwen3-0.6B and Qwen3-4B SE_{en} showing that the method is robust for native and non-native MRL models, retaining $>90\%$ of base bge-m3 performance at $k_{active} = 64$, a similar retention quality to Qwen3-0.6B and Qwen3-4B with their respective base model performances. In

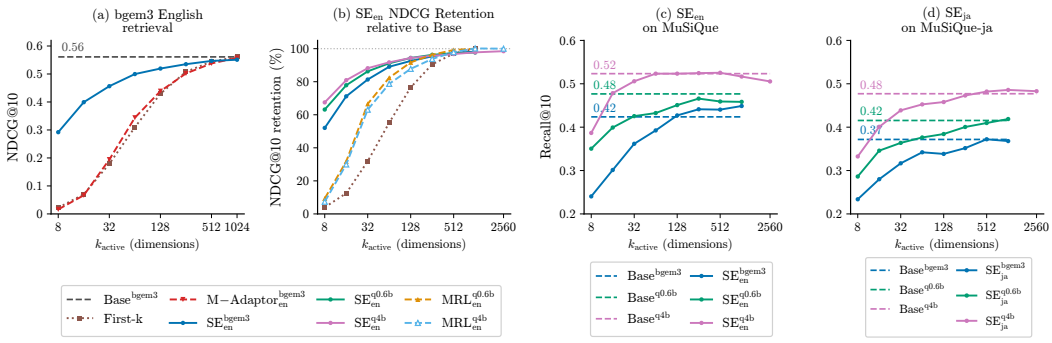


Figure 4: Performance across active embedding dimensions (k_{active}): (a) Sparsified Experts introduce ultra-sparsity in `bge-m3` and (b) NDCG@10 retention of `SE` and `MRL` are base-model-agnostic. (c–d) `SE` Recall@10 of SE$_{en}$ and SE$_{ja}$ on multi-hop datasets, ie `MuSiQue` and `MuSiQue-ja` improves performance over base model capability

complex multi-hop retrieval datasets, such as Fig. 4(c–d), SE$^{bge-m3}_{en}$ exceeds the `bge-m3` base model performance and approaches that of SE$^{q0.6b}_{en}$, which is an `MRL`-native model. Thus, *the weighted-top-k generalizes across both native and non-native `MRL` models; it helps `bge-m3`, a non-native `MRL` model, be at par with native `MRL` models in ultra-sprase representation regimes.*

5.2 Performant SEs in Ultra-sparse Regimes

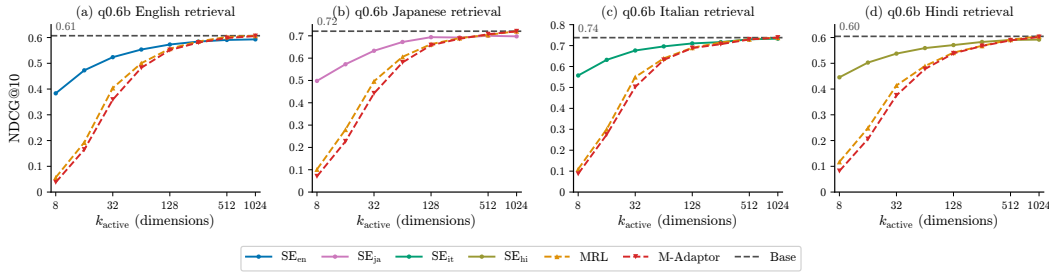


Figure 5: `Qwen3-0.6B` language experts consistently outperform `MRL` and M-Adaptor across English, Japanese, Italian, and Hindi retrieval tasks for $k_{active} \leq 128$

In Figure 1(a), we showed that SE$_{en}$, our English Sparsified Expert outperforms every baseline for $k_{active} \leq 128$. Noteably, we perform better than the methods such as `CSRv2` and `CSRv2-linear` (Guo et al., 2026) for $k_{active} > 2$, specifically designed for these ultra-sparse regimes. Due to non-reproducibility of `CSRv2` and `CSRv2-linear` experts, we directly score the authors' released retrieval-expert artifact with its released `LoRA` as the primary `CSRv2` and `CSRv2-linear` comparator. Further, our `SE`s have been trained with `mMARCO` and tested on `MMTEB`, while `CSRv2` and `CSRv2-linear` are also trained on `MMTEB` datasets, showcasing that our method is generalized better compared to these SoTA ultra-sparse methods. We also consistently outperform `Qwen3-4B` and `Qwen3-0.6B` `MRL` for $k_{MRL} \in \{32, 64, 128\}$ as shown in Figs. F.2 and 5, highlighting that our language experts, SE$_{en}$, SE$_{ja}$, SE$_{it}$, and SE$_{hi}$ outperform `Qwen3` models that are natively trained with `MRL` with $k_{MRL} \in \{32, 64, 128, ..., d\}$ (Zhang et al., 2025). We also outperform `M-Adaptor` for all $k_{MA} \in \{8, 16, 32, ..., d\}$ (full results on individual experts are in Tables F.3 to F.5). Thus, *our Sparsified Experts are `SoTA` at ultra-sparsity and beats `MRL` for* $k_{active} \leq 128$.

5.3 Task-specific Dominance of SEs and MORSE

`SE` outperforms all baselines for $k_{active} \leq 128$, after which `MRL` (i.e. a generalized embedding model) outperformed experts. We developed `MORSE` to overcome this performance gap. `MORSE` was

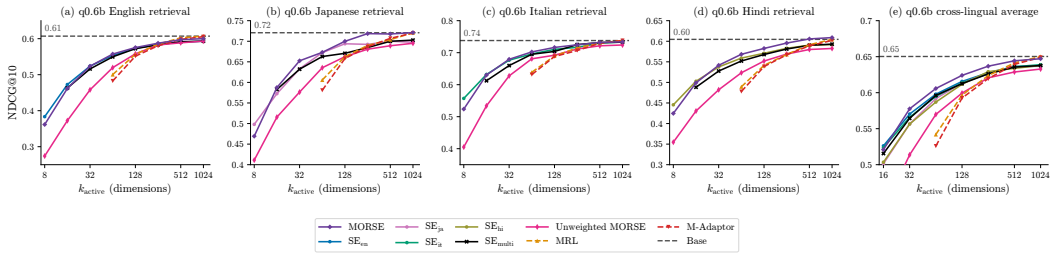


Figure 6: (a – d) Language SEs outperform MORSE at ultra-sparsity – $k_{active} \leq 32$; MORSE outperforms MRL at higher dimensions – $k_{active} \leq 512$, (e) MORSE generalizes across different languages and outperform every SE and MRL in at all dimensions

trained with a focus on $k_{active} \geq 16$, as described in Appendix C to generalize performance beyond extremely small k_{active} dimensions, and improve the overall task performance. We also trained SE_{multi}, Sparsified Expert with multiple languages, and Unweighted MORSE, a vanilla averaging across language specific SEs to baseline MORSE performance. We observe in Fig. 6 that for Qwen3-0.6B, Language specific SE outperforms MORSE in ultra-sparse regions, for $k_{active} \leq 32$, and MORSE always outperforms M-Adaptor and matches MRL and the base model's performance for all languages but English. We also see very similar trends in Qwen3-4B and bge-m3 as shown in Fig. F.2. Thus, *our proposed methods, SEs and MORSE outperform all baselines across the entire compression axis, with SE surpassing CSRv2 and CSRv2-linear at ultra-sparsity, and MORSE beating SE, M-Adaptor, and MRL at higher k_{active} dimensions while matching respective base model's performance at full dimensions.*

5.4 MORSE'S TASK-GENERALIZATION ACROSS REPRESENTATION DIMENSIONS

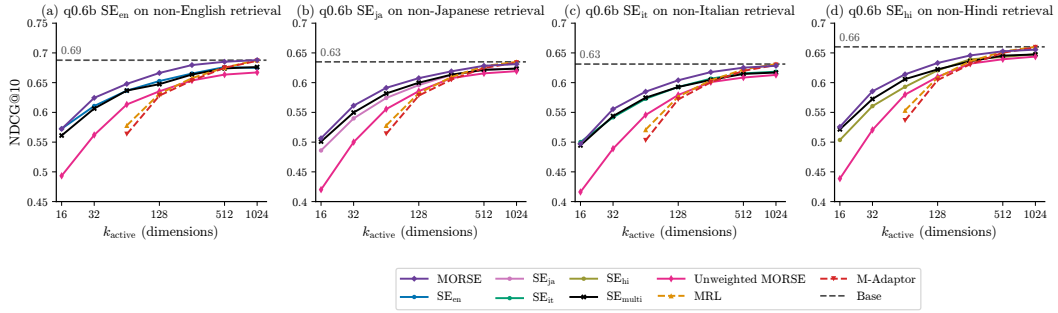


Figure 7: Cross-lingual generalization of Qwen3-0.6B: panels (a)–(d) evaluate each SE_{lang} on retrieval tasks from the other three languages against MORSE, Unweighted MORSE, SE_{multi}, MRL, M-Adaptor, and Base.

In §5.3, we saw that Language SE outperformed MORSE in very sparse dimensions for same-language retrieval tasks. In cross-lingual (i.e. task-agnostic) settings, MORSE outperforms SEs for $k_{active} \geq 16$ across models. In Figures F.3 and 7, we see that SE_{multi} actually performs on par or better than language SEs, showing multi-language training transfers to better generality. However, the Unweighted MORSE leads to poorer results than language SEs, showing that vanilla averaging is not generalizable. Finally, as expected MRL generalizes slightly better than SE on cross-lingual as compared to the same language tasks. However, MORSE continues to lead the performance consistently for $k_{active} \leq 512$, and performs similar to MRL and approaches base model's performance at full dimensions. Thus, *while there might not be an expert which outperforms all baselines from ultra-sparse dimensions to full embeddings, our method introduces a way to combine task-specific SEs that generalizes better than MRL for $k_{active} \leq 128$ and task-agnostic MORSE which outperforms every other baseline in generality, consistently outperforms MRL for $k_{active} \leq 512$, and approaches the base at full dimensions.* We delved into this interesting behavior of sparse representations in details in Appendix D.1, and show that MRL loses complete structural integrity compared to base model at ultra-sparse dimensions (leading to checkerboard and stripe patterns), while MORSE and

SEs maintain closer alignment with additional degrees of freedom. However, at $k_{active} = 1024$, native MRL overtakes MORSE in both base structure alignment and neighbor recall.

5.5 MMTEB Task Transfer

While MORSE was general across cross-lingual retrieval tasks, we also measured generalization on out-of-domain tasks in MMTEB. On eight held-out retrieval tasks (`retrieval_extended` column)), we observed that SE_{en} holds its performs compared to MRL, while CSRv2 and CSRv2-linear outperforms SE_{en} at low dimensional representations, $k_{active} < 64$ (see Figure F.4). We believe this is owing to CSRv2 and CSRv2-linear's training on the exact same datasets, while SE_{en} has only been trained entirely on mMARCO. On sevel non-retrieval task, SE_{en} performs at par with MRL for $k_{active} \leq 128$ and with CSRv2 and CSRv2-linear on the 7 non-retrieval tasks across all dimensions. Thus, *our methods generalize beyond the linguistic dimension and well across other task dimensions*.

6 Discussion

Relationship between Storage and *active* dimension k. We explain the difference between amount of storage required for MRL-based k dimensional representation and sparse representations obtained in this work or prior works (Wen et al., 2025; Guo et al., 2026). For any $k < d$, ESR or MORSE uses one fp32 and one uint16 bytes to store value in each active dimension and its index. At $k = d$, every dimension's value is stored without the need to store index.

$$\text{storage}_{\text{sparse}}(k) = \begin{cases} 6k \text{ bytes}, & k < d, \\ 4k \text{ bytes}, & k = d, \end{cases} \text{ and } \text{storage}_{\text{MRL}}(k) = 4k \text{ bytes}.$$

Thus for $k < d$, we acknowledge that ESR uses 1.5 times more storage than a MRL-based representation (*i.e.,* as it uses dense prefix) at the same k. While we report metric scores for the same k across models in the main paper, we also show that MORSE is on par or outperforms the baselines even at (approximate) same storage (see Figure F.1).

Efficiency. Each text requires one backbone pass, followed by an expert adapter and top-k selection. At $d = 1024$, an expert head uses $2.10M$ parameters (0.4% of the backbone). MORSE evaluates the four frozen expert heads after the shared backbone pass, then applies an aggregation head with $3.15M$ parameters during training and $2.10M$ at inference. This makes encoding more expensive than using one expert, but does not multiply the persistent retrieval cost *i.e.,* each document contributes one $6k$-byte representation for $k < d$ (or $4d$ bytes at the full dimension) to one shared index. Sparse matrix or inverted-index scoring can then operate directly on the k stored dimensions.

Limitations. The current objective assigns fixed weights to the supported dimensions, creating optimization imbalances across different k_{active} values and tying the resulting tradeoff to a manually chosen schedule. A next step would include learning these weights jointly with the representation, subject to constraints or regularization that prevent the objective from collapsing onto one *active* dimension. More broadly, our experiments apply this training paradigm only through post-hoc adapters. Its competitiveness in that setting motivates full-backbone training as an alternative to MRL. If implemented effectively, end-to-end training could move the current quality-storage Pareto frontier and expose new Pareto-optimal compression points, but this remains to be established experimentally.

7 Conclusion

We introduced a model-agnostic lightweight post-hoc adapter called Sparsified Expert (SE), that uses a weighted-top-k objective to create an elastic sparse representation that retains performance similarly for native and non-native MRL base models. SEs create a sparse dimensional representations while maintaining structure better than MRL as compared to respective base models. We further introduce MORSE, which consolidates independently trained experts into an elastic sparse representation improving on cross-lingual generality without sacrificing the experts' low-dimensional advantages, and performing close to base model at full dimensions. Together, SE and MORSE trace a Pareto frontier across the elastic representations, establishing sparse expert adaptation and representation-level aggregation as practical alternatives to prefix-based compression.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and editing, literature and result checks, LaTeX and plotting code, interpretation of experimental results, and refinement of the machine-translated Japanese `MuSiQue` data described in **??**. The authors reviewed all AI-assisted work and take responsibility for the final text, claims, code, data, and artifacts.

REPRODUCIBILITY STATEMENT

The main text specifies the architecture, objectives, training data, evaluation protocol, metrics, and storage accounting. Appendices A to C, E and F and **??** provide model and dataset inventories, initialization and parameter counts, training and selection details, exact external artifact identities, and supporting results. Release of code and model artifacts is pending legal approval.

REFERENCES

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL https://arxiv.org/abs/1611.09268.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset, 2021. URL https://doi.org/10.48550/arXiv.2108.13897. arXiv preprint 2021.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning Multi-Level Features with Matryoshka Sparse Autoencoders, 2025. URL https://doi.org/10.48550/arXiv.2503.17547. arXiv preprint 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024. URL https://doi.org/10.48550/arXiv.2402.03216. arXiv preprint 2024.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, et al. MMTEB: Massive Multilingual Text Embedding Benchmark, 2025. URL https://doi.org/10.48550/arXiv.2502.13595. ICLR 2025.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking, 2021. URL https://doi.org/10.48550/arXiv.2107.05720. SIGIR 2021.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized Product Quantization, 2014. URL https://doi.org/10.1109/TPAMI.2013.240. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(4):744–755.

Lixuan Guo, Yifei Wang, Tiansheng Wen, Yifan Wang, Aosong Feng, Bo Chen, Stefanie Jegelka, and Chenyu You. CSRv2: Unlocking Ultra-Sparse Embeddings, 2026. URL https://doi.org/10.48550/arXiv.2602.05735. ICLR 2026.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition, 2024. URL https://doi.org/10.48550/arXiv.2307.13269. COLM 2024.

Naamán Huerga-Pérez, Rubén Álvarez, Rubén Ferrero-Guillén, Alberto Martínez-Gutiérrez, and Javier Díez-González. Optimization of embeddings storage for RAG systems using quantization and dimensionality reduction techniques, 2025. URL https://doi.org/10.48550/arXiv.2505.00105. arXiv preprint 2025.

Phung Gia Huy, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen, and Trung Le. MIPIC: Matryoshka Representation Learning via Self-Distilled Intra-Relational and Progressive Information Chaining, 2026. URL https://doi.org/10.48550/arXiv.2604.24374. arXiv preprint 2026.

Keller Jordan. Muon: An Optimizer for Hidden Layers in Neural Networks, 2024. URL https://kellerjordan.github.io/posts/muon/. Author's technical report, December 8.

Sajjad Kachuee and Mohammad Sharifkhani. Geometry-Preserving Aggregation for Mixture-of-Experts Embedding Models, 2026. URL https://doi.org/10.48550/arXiv.2602.14039. arXiv preprint 2026.

Pranav Kasela, Gabriella Pasi, Raffaele Perego, and Nicola Tonellotto. DESIRE-ME: Domain-Enhanced Supervised Information REtrieval using Mixture-of-Experts, 2024. URL https://doi.org/10.48550/arXiv.2403.13468. ECIR 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2015. URL https://doi.org/10.48550/arXiv.1412.6980. ICLR 2015.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, et al. Matryoshka Representation Learning, 2022. URL https://doi.org/10.48550/arXiv.2205.13147. NeurIPS 2022.

Hyunji Lee, Luca Soldaini, Arman Cohan, Minjoon Seo, and Kyle Lo. RouterRetriever: Routing over a Mixture of Expert Embedding Models, 2025. URL https://doi.org/10.48550/arXiv.2409.02685. AAAI 2025.

Alireza Makhzani and Brendan Frey. k-Sparse Autoencoders, 2014. URL https://doi.org/10.48550/arXiv.1312.5663. ICLR 2014.

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to Route Among Specialized Experts for Zero-Shot Generalization, 2024. URL https://doi.org/10.48550/arXiv.2402.05859. ICML 2024.

Zach Nussbaum and Brandon Duderstadt. Training Sparse Mixture Of Experts Text Embedding Models, 2025. URL https://doi.org/10.48550/arXiv.2502.07972. arXiv preprint 2025.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond, 2009. URL https://doi.org/10.1561/1500000019. Foundations and Trends in Information Retrieval.

Miguel Romero, Shuoyang Ding, Corey D. Barret, Georgiana Dinu, and George Karypis. Beyond instruction-conditioning, MoTE: Mixture of Task Experts for Multi-task Embedding Models, 2025. URL https://doi.org/10.48550/arXiv.2506.17781. arXiv preprint 2025.

Effrosyni Sokli, Pranav Kasela, Georgios Peikos, and Gabriella Pasi. Investigating Mixture of Experts in Dense Retrieval, 2024. URL https://doi.org/10.48550/arXiv.2412.11864. arXiv preprint 2024.

Sotaro Takeshita, Yurina Takeshita, Simone Paolo Ponzetto, and Daniel Ruffinelli. To MRL or not to MRL: Text Embeddings are Robust to Truncation Without Matryoshka Learning, Except In Heavy Truncation Scenarios, 2026. URL https://doi.org/10.48550/arXiv.2605.16608. arXiv preprint 2026.

Hayato Tsukagoshi and Ryohei Sasano. Redundancy, Isotropy, and Intrinsic Dimensionality of Prompt-based Text Embeddings, 2025. URL https://doi.org/10.48550/arXiv.2506.01435. arXiv preprint 2025.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond Matryoshka: Revisiting Sparse Coding for Adaptive Representation, 2025. URL https://doi.org/10.48550/arXiv.2503.01776. ICML 2025.

Jinsung Yoon, Rajarishi Sinha, Sercan O Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10318–10336, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.576. URL https://aclanthology.org/2024.emnlp-main.576/.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate, 2025. URL https://doi.org/10.48550/arXiv.2504.19874. arXiv preprint 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, et al. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, 2025. URL https://doi.org/10.48550/arXiv.2506.05176. arXiv preprint 2025.

The appendix provides the experimental records and supporting results. Appendices A to C specify models, datasets, initialization, and training. Appendices E and F and **??** report benchmark specialization, detailed scores, and language analyses.

A MODELS AND DATASETS

evaluation	language	datasets or task categories
english.retrieval	english	BelebeleRetrieval.eng_Latn-eng_Latn, MultilingualNanoNQRetrieval.en, MultilingualNanoHotpotQARetrieval.en, MultilingualNanoFEVERRetrieval.en, MultilingualNanoDBPediaRetrieval.en.
japanese.retrieval	japanese	BelebeleRetrieval.jpn_Jpan-jpn_Jpan, MultilingualNanoNQRetrieval.ja, MultilingualNanoHotpotQARetrieval.ja, MultilingualNanoFEVERRetrieval.ja, MultilingualNanoDBPediaRetrieval.ja.
italian.retrieval	italian	BelebeleRetrieval.ita_Latn-ita_Latn, MultilingualNanoNQRetrieval.it, MultilingualNanoHotpotQARetrieval.it, MultilingualNanoFEVERRetrieval.it, MultilingualNanoDBPediaRetrieval.it.
hindi.retrieval	hindi	BelebeleRetrieval.hin_Deva-hin_Deva, MLQARetrieval.hin-hin, MintakaRetrieval.hi, XPQARetrieval.hin-hin, IndicQARetrieval.hi.
english.retrieval_extended	english	FiQA2018.default, NFCorpus.default, SciFact.default, ArguAna.default, Climate FEVERHardNegatives.default, CQADupstack GamingRetrieval.default, CQADupstackUnix Retrieval.default, SCIDOCS.default.
english.retrieval_hard	english	MuSiQue.
japanese.retrieval_hard	japanese	musique_ja.
english.tasks	english	BibleNLPBitextMining.eng_Latn-aai_Latn (bitextmining), EmotionClassification.v2.default (classification), ArXivHierarchical ClusteringP2P.default (clustering), RTE3.en (pairclassification), CorebC2TReranking.default (reranking), BIOSSES.default (sts), SummEval Summarization.v2.default (summarization).

Table A.1: Evaluation inventory. Dataset names include the evaluated subset where applicable, and task categories appear in parentheses.

backbone	architecture	params.	d	native
BGE-M3	encoder-only	568M	1024	None
Qwen3-Embedding-0.6B	decoder-only	0.6B	1024	MRL
Qwen3-Embedding-4B	decoder-only	4B	2560	MRL

Table A.2: Evaluated backbones and their native compression support. All backbones emit normalized fp32 embeddings.

backbone	adapters
BGE-M3	english_expert, japanese_expert, italian_expert, hindi_expert.
Qwen3-Embedding-0.6B	english_expert, japanese_expert, italian_expert, hindi_expert, multilingual_expert, morse.
Qwen3-Embedding-4B	english_expert, english_mteb_expert.

Table A.3: Adapter inventory by backbone.

B ARCHITECTURE DETAILS

Expert initialization and size. Both encoder matrices are initialized orthogonally, while b_{center} and b_{latent} are initialized at zero. The expert uses $2(d^2 + d)$ parameters during both training and inference because the tied decoder introduces no additional parameters. This is 2.10 million at $d = 1024$ and 13.11 million at $d = 2560$. Only the encoder and top-k selection are required at inference.

MORSE initialization and size. The projections P_j are initialized orthogonally. In the reported setting, $W^{(\text{encode})}$ and $W^{(\text{decode})}$ are initialized as identity matrices, and all biases are initialized at zero. The aggregation head uses $8pd + 4p + d$ parameters at inference and $d^2 + 8pd + 4p + 2d$ during training. For $d = 1024$ and $p = 256$, these counts are 2.10 million at inference and 3.15 million during training, in addition to the four frozen expert heads.

Relation to sparse coding. CSR and CSRv2 motivate sparse post-hoc specialization, but the reported main experts keep the backbone frozen and learn a signed latent space of dimension d, without an expanded dictionary. A single weighted-top-k objective supports several storage budgets. The second stage consolidates compatible experts into one stored code, rather than selecting a separate document index at query time. These distinctions concern the representation, supervision, and deployment setting, rather than sparse coding alone.

Prefix sizes. Qwen3-Embedding supports prefix sizes from 32 to the full dimension, so intermediate sizes such as mrl_48, mrl_96, and mrl_192 are valid prefix evaluations. The discrete prefix sizes used in training are not listed in the technical report (Zhang et al., 2025).

Additional context. For prompt-conditioned embeddings, robustness to dimensionality reduction varies by task and correlates with isotropy and intrinsic dimensionality (Tsukagoshi & Sasano, 2025). Training-time extensions such as MIPIC strengthen low-dimensional prefixes through relational self-distillation and progressive information chaining (Huy et al., 2026). A post-training comparison evaluates quantization and dimensionality reduction through explicit storage-quality trade-offs (Huerga-Pérez et al., 2025). Separate work on internal model activations, rather than final embeddings, trains Matryoshka SAEs with nested dictionary prefixes to retain interpretable features at multiple levels of abstraction (Bussmann et al., 2025). DESIRE-ME combines query-side domain specialists after a dense encoder (Kasela et al., 2024), while SB-MoE applies one trainable MoE block to final query and document embeddings (Sokli et al., 2024). Nomic Embed v2 instead places sparse expert layers inside the encoder (Nussbaum & Duderstadt, 2025). Within sparse MoE encoders, spherical barycentric aggregation replaces linear mixing of routed expert outputs with a geometry-aware operator (Kachuee & Sharifkhani, 2026). Outside embedding models, LoraHub learns task-level coefficients over frozen LoRA modules (Huang et al., 2024), while PHATGOOSE learns post-hoc, input-dependent gates over specialized adapters (Muqeeth et al., 2024).

C IMPLEMENTATION DETAILS

The main experts train for 10 epochs on 100,000 triples per language. A separate 20,000-triple split per language is reserved for validation. Training uses batch size 512 and seed 0. The multilingual expert uses the same language-expert architecture and optimization. Its training split contains 25,000 triples per language, and its separate validation split contains 5,000 per language. Each 512-row training batch contains 128 rows per language. The `MTEB`-trained expert trains on 100,000 triples across the eight `CSRv2` retrieval datasets, with a separate 20,000-triple validation split. The `MTEB`-trained expert is produced by five checkpoint-linked training stages lasting 10, 5, 2, 1, and 8 epochs. Each stage uses the same architecture and dataset, initializes from its predecessor's validation-selected checkpoint, and rebuilds `Muon`, `Adam`, the schedules, and local epoch seeding. `MORSE` trains for 10 epochs on 25,000 triples per language, with a separate 5,000-triple validation split per language, batch size 128, and seed 0. Each optimization stage uses linear warmup over the first 10% of its steps, cosine decay to 10% of its peak learning rate, and gradient clipping at norm 1. Expert checkpoints are selected by validation `recall@10`, while `MORSE` is selected by validation `nDCG@10` at $k = 32$.

Experts. The regularization coefficients are $\lambda_{\text{contrastive}} = 0.25$ and $\lambda_{\text{distillation}} = 3$, with temperature $\tau = 0.05$ and `KL` distillation. `Muon` uses learning rate 0.03, momentum 0.95, and weight decay 0.01 for matrix parameters. `Adam` uses learning rate 0.0012, $\beta = (0.9, 0.95)$, $\epsilon = 6.25 \times 10^{-10}$, and weight decay 0.01 for vector parameters. For the 1024-dimensional backbones:

$$\mathcal{K} = \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\},$$
$$(w_2, \ldots, w_{1024}) \propto (0.02, 0.04, 0.08, 0.4, 0.3, 0.08, 0.04, 0.02, 0.01, 0.01).$$

The `Qwen3-Embedding-4B` expert uses:

model	weights	distillation	primary	extended
`linear SAE`	`uniform`	`None`	0.6279 / 0.7125	0.2910 / 0.3642
`linear SAE`	`uniform`	ℓ_2	0.6493 / 0.7348	0.3022 / 0.3793
`linear SAE`	`weighted`	ℓ_2	0.6655 / 0.7249	0.3257 / 0.3850
`deep SAE`	`weighted`	ℓ_2	0.6789 / 0.7174	0.3297 / 0.3889
`deep SAE`	`weighted`	D_{KL}	0.6872 / 0.7311	0.3312 / 0.3909

Table C.1: Component ablation for the `BGE-M3 english` expert. Each result cell reports mean `recall@10` at $k = 16$ / $k = 32$. Primary and extended retrieval refer to the five- and eight-dataset English groups in Table A.1. Every row is trained only on `mMARCO`.

$$\mathcal{K} = \{2, 4, 8, 16, 32, 64, 128, 256, 512, 640, 1024, 2560\},$$
$$(w_2, \ldots, w_{2560}) \propto (0.02, 0.05, 0.08, 0.35, 0.25, 0.08, 0.08, 0.03, 0.02, 0.02, 0.01, 0.01).$$

The `MTEB`-trained expert instead uses:

$$\mathcal{K} = \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 2560\},$$
$$(w_2, \ldots, w_{2560}) \propto (0.02, 0.03, 0.2, 0.12, 0.15, 0.15, 0.12, 0.09, 0.06, 0.03, 0.015, 0.015).$$

The final `MTEB`-trained stage weights reconstruction, contrastive retrieval, and distillation by 1.1, 0.1, and 10.0, respectively, with $k_{\text{selection}} = 8$ and $\tau = 0.075$. The reconstruction weight is 1.0 in the first four stages. Its `Muon` and `Adam` learning rates are 1.875×10^{-3} and 7.5×10^{-5}, respectively. Stage 1 uses the initial weight schedule $(0.02, 0.03, 0.07, 0.14, 0.18, 0.18, 0.15, 0.11, 0.06, 0.03, 0.015, 0.015)$. Stages 2 through 5 use the `MTEB` weights above. The stages select local epochs 9, 5, 2, 1, and 8, respectively. The final selected state attains validation `recall@10` 0.7414.

With backbone embeddings cached, the reported 10-epoch expert training takes < 5 minutes on a single NVIDIA `L40S` GPU for both `BGE-M3` and `Qwen3-Embedding`. On a saved `SciFact` probe with 100 queries and 5183 documents, the `BGE-M3` English expert at $k = 16$ takes $1.027\times$ the dense base's total online time, a 2.65% overhead including backbone encoding, the sparse head, and scoring.

`MORSE`. The alignment coefficient is $\lambda_{\text{alignment}} = 0.25$. `MORSE` uses the main experts' `Muon` settings. `Adam` uses learning rate 0.0006, $\beta = (0.9, 0.95)$, $\epsilon = 6.25 \times 10^{-10}$, and weight decay 0.01 for vector

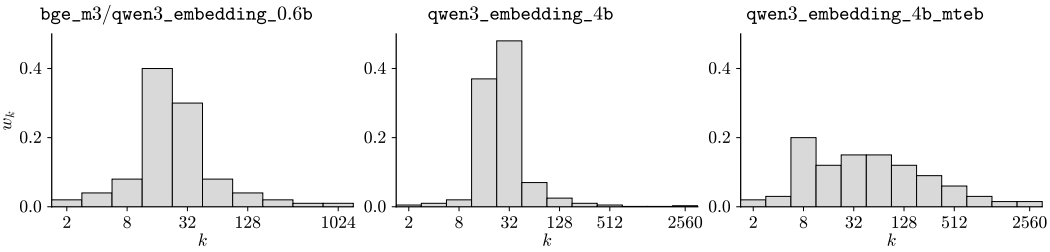


Figure C.1: weighted-top-k training-weight distributions for the main expert recipe and the two `Qwen3-Embedding-4B` recipes.

stage	epochs	schedule	$k_{\text{selection}}$	$\lambda_{\text{contrastive}}$	$\lambda_{\text{distillation}}$	τ	lr (`Muon`, `Adam`)
1.	10	initial	16	0	3.5	0.05	$(0.03, 0.0012)$
2.	5	mteb	8	0.02	10	0.05	$(3.75 \times 10^{-3}, 1.5 \times 10^{-4})$
3.	2	mteb	8	0.05	3.5	0.075	$(1.875 \times 10^{-3}, 7.5 \times 10^{-5})$
4.	1	mteb	8	0.05	3.5	0.075	$(1.875 \times 10^{-3}, 7.5 \times 10^{-5})$
5.	8	mteb	8	0.1	10.0	0.075	$(1.875 \times 10^{-3}, 7.5 \times 10^{-5})$

Table C.2: Complete checkpoint-linked training chain for the `MTEB`-trained expert. Each stage initializes from its predecessor's validation-selected checkpoint. The schedule column refers to the weights over active dimensions defined above. The k column gives the checkpoint-selection dimension. Loss coefficients are the implemented absolute weights, with reconstruction weighted by 1.0 in stages 1–4 and 1.1 in stage 5.

parameters. The encoding map $\left(W^{(\text{encode})}, b^{(\text{encode})}\right)$ uses a learning-rate multiplier of 0.0333. Each input expert remains at $k_{\text{input}} = 1024$ throughout training, and the projection dimension is $p = 256$. The output objective starts with all weight on $k = 1024$ and, over the first half of training, transitions to uniform weights over $\mathcal{K}_{\text{MORSE}} = \{16, 32, 64, 128, 256, 512, 1024\}$.

D Interpretability Study

D.1 Effect of Sparsity on Structural Geometry

Setup. We study how specialization and expert composition change the geometry of compressed query representations. The experiment uses 900 English Belebele queries encoded by Qwen3-Embedding-4B. Let $z_{\text{B}}(x) \in \mathbb{R}^{2560}$ denote the normalized Base representation and let $T_k(\cdot)$ retain the k coordinates with largest absolute magnitude. At each budget

$$k \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 2560\},$$

we compare four representations:

$$z^{(k)}_{\text{MRL}}(x) = \text{norm}(z_{\text{B}}(x)_{1:k}), \tag{7}$$

$$z^{(k)}_{\text{Expert}}(x) = \text{norm}(T_k(e_{\text{en}}(x))), \tag{8}$$

$$z^{(k)}_{\text{MORSE}}(x) = \text{norm}(T_k(m(x))). \tag{9}$$

Here e_{en} is the English specialist and m is the trained MORSE mixer over the language experts. Native MRL uses the first k Base coordinates, whereas the expert and MORSE select an input-dependent support. The full Base representation is included as the uncompressed reference.

For row-normalized representations $Z^{(k)}_s \in \mathbb{R}^{n \times d_s}$, we compare the cosine Gram matrices

$$G^{(k)}_s = Z^{(k)}_s Z^{(k)\top}_s.$$

Figure D.1 displays the same 96 queries in every panel. Their order is fixed by clustering the Base Gram matrix and then reused across systems and widths. Every panel uses the same cosine range $[-1, 1]$, so color differences are directly comparable.

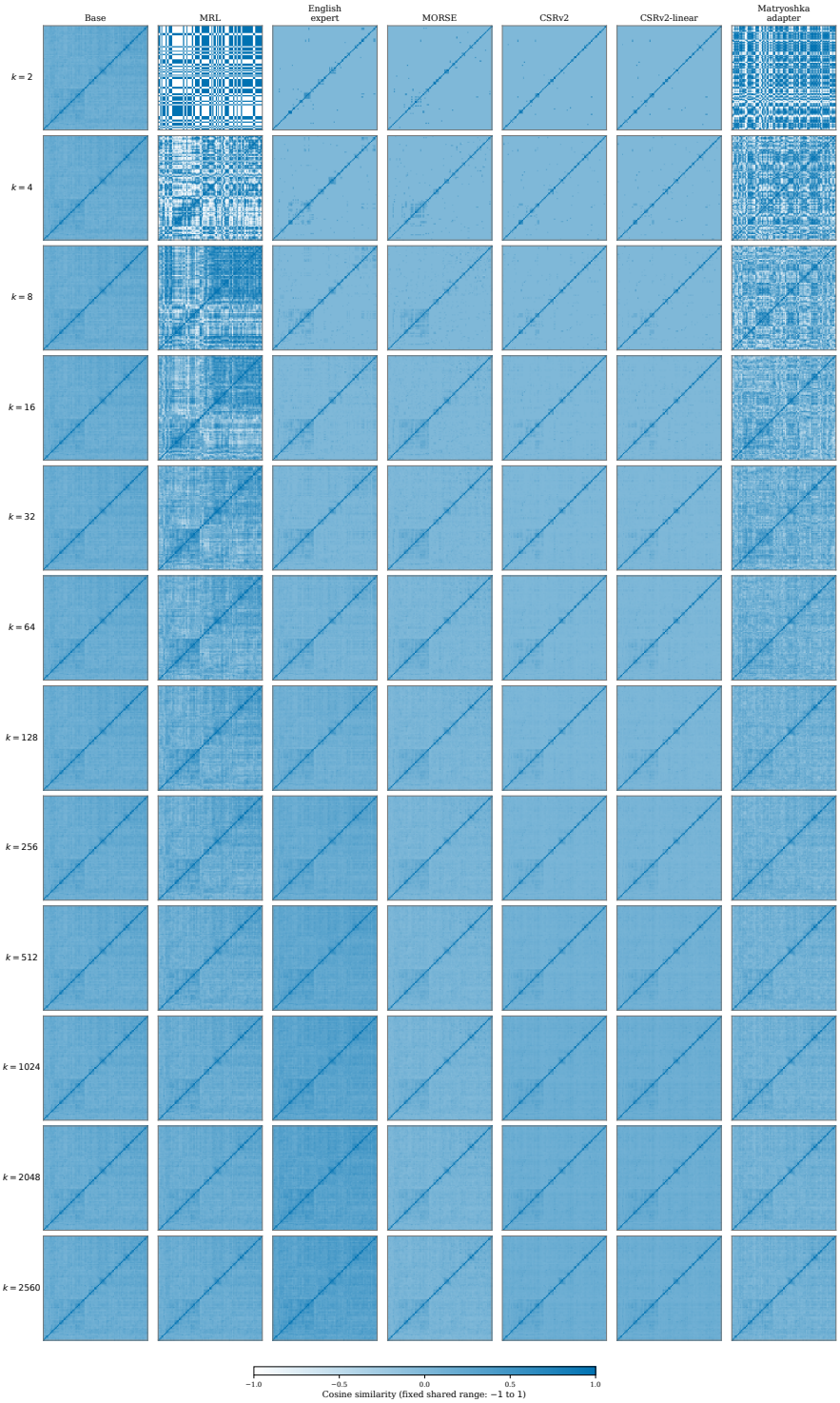


Figure D.1: Base-ordered cosine-similarity matrices for Qwen3-Embedding-4B. Widths are arranged in two six-row blocks to reduce vertical space; columns compare Base, native MRL, the English expert, and trained MORSE. At small k, MRL exhibits pronounced shared-prefix artifacts, while the expert and MORSE retain more of the Base neighborhood structure.

Geometry measures. Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. We quantify centered relational alignment with

$$\mathrm{CKA}_s(k) = \frac{\left\langle HG_s^{(k)}H, HG_{\mathrm{B}}H \right\rangle_F}{\left\| HG_s^{(k)}H \right\|_F \left\| HG_{\mathrm{B}}H \right\|_F}.$$

We also report Base-neighbor recall@10, the mean fraction of each query's ten nearest Base neighbors recovered by the compressed representation. Base has value one for both measures.

Key findings. The low-width difference has a direct geometric explanation. Native MRL uses one common k-coordinate prefix, and therefore

$$\mathrm{rank}\left(G_{\mathrm{MRL}}^{(k)}\right) = \mathrm{rank}\left(Z_{\mathrm{MRL}}^{(k)}\right) \le k.$$

The English expert and MORSE instead use query-dependent supports S_i. Their sample rank is bounded by the union of those supports,

$$\mathrm{rank}\left(G_s^{(k)}\right) \le \min\left(n, \left|\bigcup_i S_i\right|\right),$$

and need not be bounded by the per-query budget k. At $k = 32$, native MRL has measured numerical rank 30, while both the English expert and MORSE have rank 899 on the 900-query sample. This removes the global shared-subspace bottleneck visible as checkerboard and stripe patterns in the MRL heatmaps.

The alignment measures show that the additional degrees of freedom preserve Base-relevant relations. At $k = 32$, the English expert and MORSE obtain CKA 0.856 and 0.826, compared with 0.640 for MRL. Their Base-neighbor recalls are 0.632 and 0.605, compared with 0.382 for MRL. The English expert remains strongest among the compressed systems shown through $k = 512$, with MORSE consistently second. Across $k \le 512$, MORSE remains within 0.040 CKA and 0.041 neighbor recall of the English specialist. Thus, the learned MORSE mixture retains most of the specialist's relational geometry while composing multiple language experts.

This advantage is specific to the compressed regime. At $k = 1024$, native MRL overtakes MORSE in both CKA and neighbor recall. At $k = 2560$, the MRL prefix contains every Base coordinate, giving

$$G_{\mathrm{MRL}}^{(2560)} = G_{\mathrm{B}}$$

by construction. The English expert and MORSE remain learned transformations, so using all of their output coordinates does not imply equality with Base. The evidence therefore supports a budget-dependent conclusion: query-dependent expert supports preserve Base-relative geometry more efficiently at small and moderate active budgets, whereas a sufficiently wide native prefix directly recovers the Base representation.

D.2 Language-Expert Collision Analysis

We analyze how trained MORSE combines four language experts—English, Japanese, Italian, and Hindi—in a fixed Qwen3-Embedding-0.6B checkpoint. The evaluation uses translation-aligned Belebele retrieval subsets containing 900 queries and 488 corpus items per language, with $k \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$ retained atoms. Let $c_{e,j}(x)$ denote expert e's signed contribution to retained atom j for input x. We define its absolute ownership share and the resulting collision mass as

$$p_{e,j}(x) = \frac{|c_{e,j}(x)|}{\sum_f |c_{f,j}(x)|}, \qquad C_j(x) = 1 - \sum_e p_{e,j}(x)^2. \tag{10}$$

For four experts, $0 \le C_j \le 3/4$; its inverse-Simpson transform $n_{2,j} = 1/(1 - C_j)$ is the effective number of experts contributing to the atom. Thus, low collision means concentrated expert ownership, but does *not* by itself imply lower representation rank, lower intrinsic dimension, or agreement among the signed contributions.

Training produces substantially more concentrated ownership than uniform expert aggregation (Figure D.2a). The trained-minus-unweighted collision difference is negative for every one of the 32

language-by-width query comparisons and remains significant after Holm correction; at $k = 32$, the paired differences are -0.218, -0.225, -0.222, and -0.230 for English, Japanese, Italian, and Hindi, respectively. Averaged across languages, trained collision rises from 0.271 at $k = 8$ to 0.608 at $k = 1024$, while the corresponding advantage over uniform aggregation narrows from 0.297 to 0.025. The effective expert count consequently grows from $n_2 = 1.43$ to 2.72: the highest-magnitude atoms are primarily owned by one expert, whereas lower-magnitude atoms admitted at larger k are increasingly shared. Signed cancellation also rises from 0.042 to 0.621, so concentrated ownership should not be conflated with expert agreement.

The learned concentration is a shared specialization pattern rather than a hard language router (Figure D.2b). At $k = 32$, English inputs allocate mean absolute contribution shares $(0.354, 0.308, 0.195, 0.144)$ to the English, Japanese, Italian, and Hindi experts, respectively; Japanese, Italian, and Hindi inputs exhibit the same ordering. Moreover, the matching expert is top-ranked for only 66.2%, 32.7%, 5.8%, and 0.7% of English, Japanese, Italian, and Hindi queries.

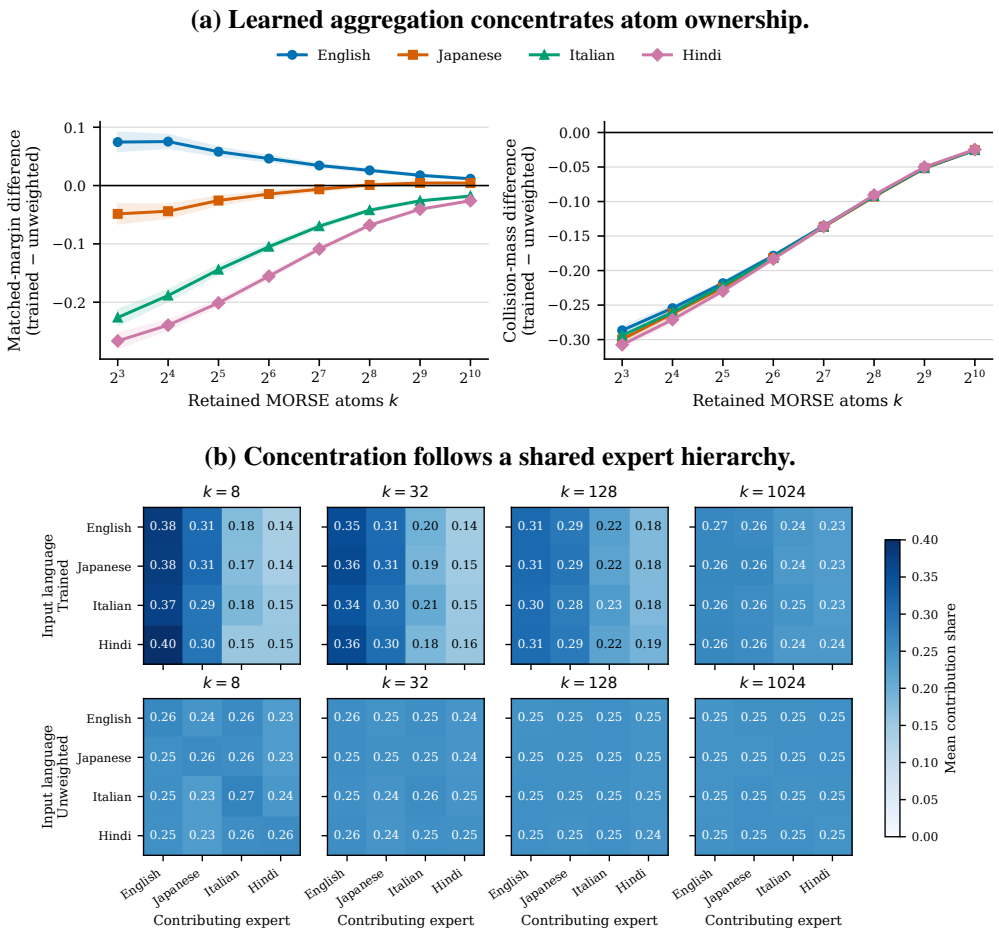


Figure D.2: Language-expert collision and contribution structure. **(a)** Trained-minus-unweighted effects on query representations; negative collision-mass differences indicate more concentrated ownership under trained MORSE, with shaded 95% paired-bootstrap confidence intervals. **(b)** Mean absolute contribution shares by input language and expert. Trained MORSE (top) learns a stable hierarchy at small and moderate k, while equal aggregation (bottom) is nearly uniform; similar rows across languages rule out a hard matching-language switch.

E CSRv2 COMPARISON

All CSRv2 evaluations use the authors' released retrieval artifact, CSRv2-retrieval at revision af16198. Its latent dictionary has 10240 coordinates, so the plotted $k = 2560$ codes still require sparse coordinate indices. Figure E.1 contrasts general retrieval supervision with the benchmark-specialized setting of CSRv2 (Guo et al., 2026), which we reproduce in our pipeline with an additional MTEB-trained expert. At $k = 32$, this diagnostic reaches 0.6433 mean recall on the eight specialized datasets, compared with 0.5897 for the released CSRv2 model and 0.5880 for its LoRA-disabled ablation. On the primary English retrieval suite, the mMARCO expert instead reaches 0.7765 vs. 0.7351 for the diagnostic, supporting a specialization-transfer tradeoff under the compared recipes. The recipes differ in optimization and total epochs as well as data (Appendix C).

Benchmark-training limitation. Benchmark-specific supervision, as explored by CSRv2, can introduce contamination risks when native training data are unavailable. In our reproduction, five sources use internal partitions of their native test data for training, so their specialized scores measure benchmark fit rather than held-out generalization. The main mMARCO experts do not train on these benchmarks.

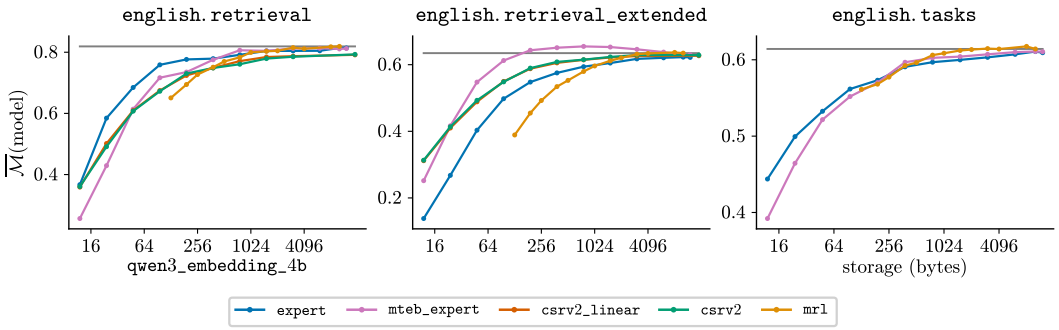


Figure E.1: Qwen3-Embedding-4B performance against per-vector storage. The panels compare the mMARCO expert, the MTEB-trained diagnostic, the released CSRv2 model and its LoRA-disabled ablation, and native MRL on the primary English retrieval suite, the eight specialized retrieval datasets, and seven non-retrieval tasks. Retrieval panels report mean recall@10, while the task panel reports the mean primary metric across the seven non-retrieval tasks. Gray lines mark dense-backbone scores.

F DETAILED RESULTS

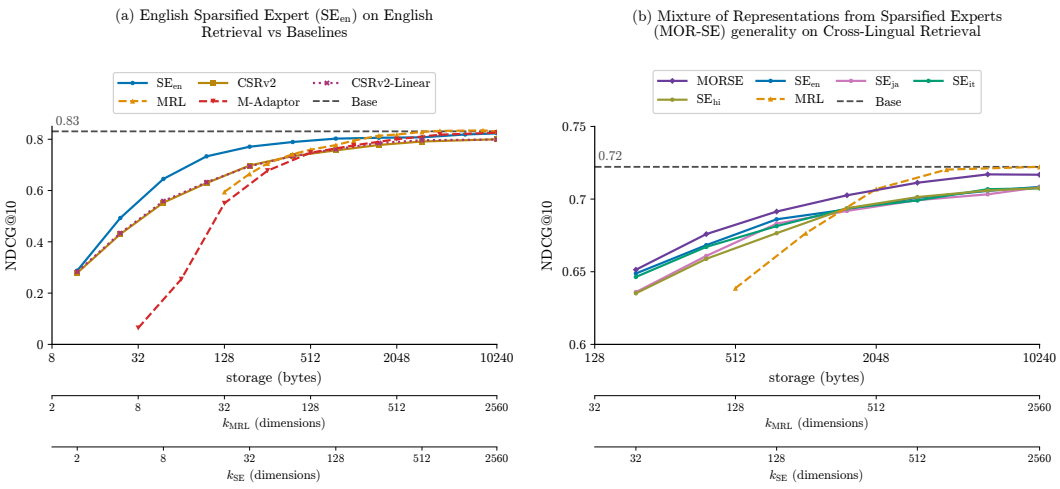


Figure F.1: (a) compares SE_{en} with compression baselines on English retrieval, while (b) compares MORSE and language experts using unweighted cross-lingual averages for q4b. Storage is the primary x-axis; k_{MRL} and k_{SE} show the dimensions achievable under the corresponding storage budget.

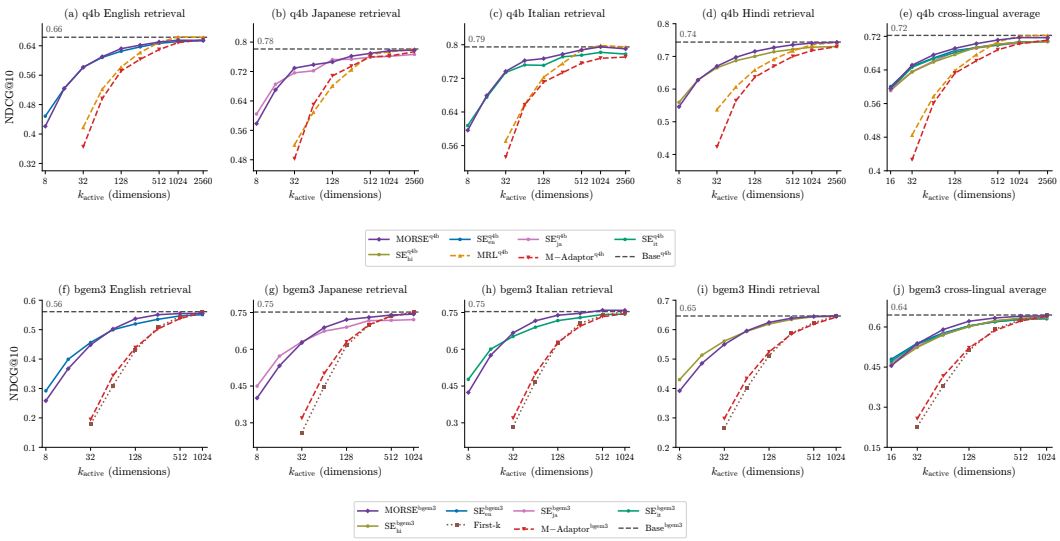


Figure F.2: Retrieval performance of `Qwen3-4B`(top) and `bge-m3`(bottom) across four language-specific tasks and the 28-dataset cross-lingual average. Each row compares MORSE and sparsified experts against compression baselines, with First-k replacing MRL for bgem3.

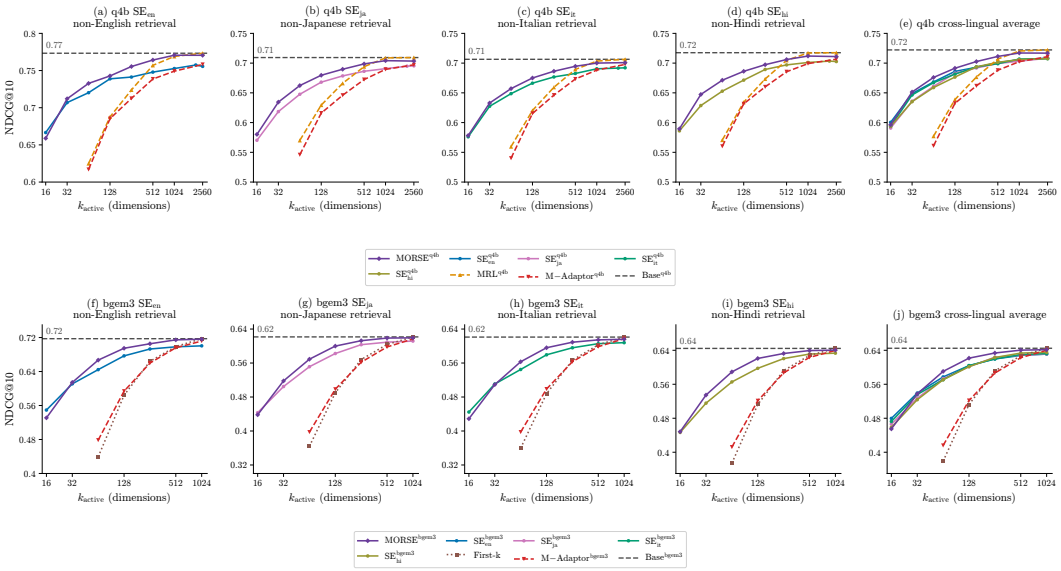


Figure F.3: Cross-lingual retrieval performance of q4b (top) and bgem3 (bottom) using language-specific experts evaluated outside their training language. Panels (e) and (j) report the all-language average; First-k replaces MRL for bgem3.

	$k = 16$		$k = 32$		$k = 48$
dataset	**expert**	**CSRv2**	**expert**	**CSRv2**	**MRL**
`recall@10`					
`Belebele`	0.9644	0.9344	0.9767	0.9622	0.9456
`NanoNQ`	0.7600	0.6600	0.7900	0.7500	0.6700
`NanoHotpotQA`	0.7900	0.5600	0.8100	0.6900	0.6800
`NanoFEVER`	0.9667	0.9267	0.9733	0.9633	0.9067
`NanoDBPedia`	0.3129	0.2807	0.3325	0.2857	0.2701
mean	0.7588	0.6724	0.7765	0.7302	0.6945
`nDCG@10`					
`Belebele`	0.9127	0.8586	0.9310	0.8942	0.8805
`NanoNQ`	0.5921	0.5453	0.6218	0.5924	0.5291
`NanoHotpotQA`	0.7207	0.5065	0.7764	0.6317	0.6030
`NanoFEVER`	0.8674	0.7447	0.9084	0.8357	0.7927
`NanoDBPedia`	0.5738	0.4898	0.6168	0.5340	0.5201
mean	0.7333	0.6290	0.7709	0.6976	0.6651

Table F.1: Per-dataset English retrieval for `Qwen3-Embedding-4B`. Expert and `CSRv2` columns at the same k are storage-matched, as are their $k = 32$ codes and `MRL` at $k = 48$. Dataset subsets are listed in Table A.1.

BGE-M3									
task	**metric**	$\texttt{SE}_{32}$	**Top-k**$_{32}$	$\texttt{SE}_{64}$	**Top-k**$_{64}$	$\texttt{SE}_{128}$	**Top-k**$_{128}$	$\texttt{SE}_{256}$	**Top-k**$_{256}$
bitext mining	F1	0.0769	0.0207	0.0865	0.0573	0.0929	0.0795	0.1082	0.1046
classification	accuracy	0.4301	0.3731	0.4481	0.4096	0.4638	0.4440	0.4681	0.4596
clustering	V-measure	0.5232	0.4935	0.5175	0.5068	0.5217	0.5051	0.5242	0.5227
pair classification	max AP	0.8768	0.8796	0.8812	0.8835	0.8788	0.8870	0.8800	0.8845
reranking	MAP@10	0.6209	0.4154	0.6404	0.5620	0.6785	0.6263	0.6929	0.6973
STS	Spearman	0.7557	0.6850	0.8307	0.7768	0.8374	0.8041	0.8365	0.8268
summarization	Spearman	0.2917	0.2531	0.3196	0.3279	0.3398	0.3216	0.3435	0.3399
mean		0.5107	0.4458	0.5320	0.5034	0.5447	0.5239	0.5505	0.5479

Qwen3-Embedding-0.6B									
task	**metric**	$\texttt{SE}_{32}$	**MRL**$_{48}$	$\texttt{SE}_{64}$	**MRL**$_{96}$	$\texttt{SE}_{128}$	**MRL**$_{192}$	$\texttt{SE}_{256}$	**MRL**$_{384}$
bitext mining	F1	0.0685	0.0547	0.0692	0.0572	0.0772	0.0703	0.0946	0.0829
classification	accuracy	0.5950	0.5898	0.6052	0.6149	0.6140	0.6199	0.6220	0.6212
clustering	V-measure	0.5410	0.5651	0.5463	0.5730	0.5499	0.5917	0.5586	0.5759
pair classification	max AP	0.8711	0.8724	0.8728	0.8748	0.8709	0.8777	0.8718	0.8806
reranking	MAP@10	0.8517	0.8210	0.8619	0.8446	0.8736	0.8696	0.8899	0.8684
STS	Spearman	0.7497	0.8012	0.7956	0.8465	0.8219	0.8536	0.8367	0.8617
summarization	Spearman	0.1986	0.2494	0.2488	0.2787	0.2582	0.2818	0.2752	0.2650
mean		0.5537	0.5648	0.5714	0.5842	0.5808	0.5950	0.5927	0.5937

Qwen3-Embedding-4B									
task	**metric**	$\texttt{SE}_{32}$	**MRL**$_{48}$	$\texttt{SE}_{64}$	**MRL**$_{96}$	$\texttt{SE}_{128}$	**MRL**$_{192}$	$\texttt{SE}_{256}$	**MRL**$_{384}$
bitext mining	F1	0.0786	0.0496	0.1127	0.0584	0.0953	0.0898	0.0975	0.0970
classification	accuracy	0.6621	0.6527	0.6712	0.6669	0.6736	0.6746	0.6730	0.6736
clustering	V-measure	0.5393	0.5284	0.5348	0.5485	0.5360	0.5680	0.5389	0.5815
pair classification	max AP	0.8581	0.8825	0.8699	0.8851	0.8751	0.8824	0.8763	0.8831
reranking	MAP@10	0.8789	0.8503	0.8993	0.8797	0.9071	0.9052	0.9116	0.8937
STS	Spearman	0.7512	0.7170	0.7838	0.7868	0.7954	0.8079	0.8043	0.8267
summarization	Spearman	0.2441	0.2958	0.2636	0.3210	0.2955	0.3148	0.2976	0.3294
mean		0.5732	0.5681	0.5908	0.5924	0.5969	0.6061	0.5999	0.6121

Table F.2: English non-retrieval task scores for $\texttt{SE}$ at $k \in \{32, 64, 128, 256\}$ with storage-matched native baselines. BGE-M3 uses Top-k at the same k, while the Qwen models use MRL at $1.5k$. The mean is a macro summary of heterogeneous primary metrics, not a full MMTEB score.

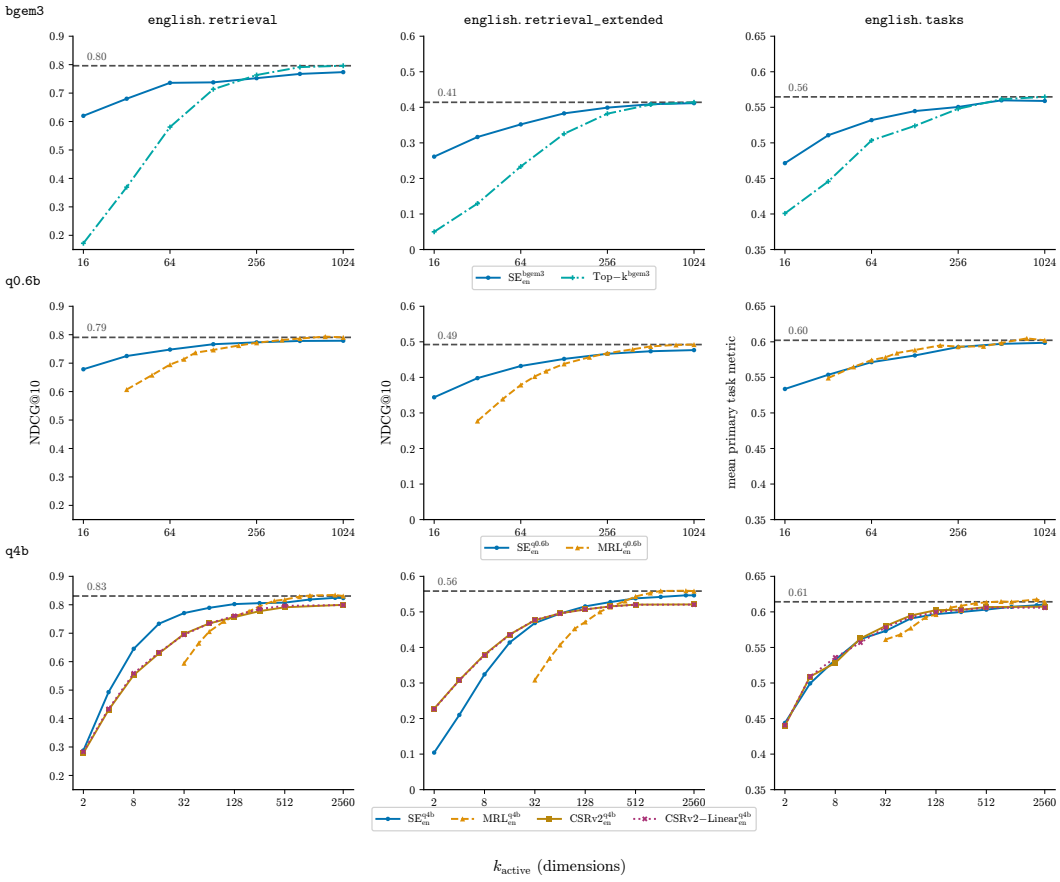


Figure F.4: Transfer beyond `mMARCO` to `MMTEB` extended retreival tasks on 8 datasets and 7 non-retrieval tasks mentioned in Table F.2

Table F.3: MMTEB retrieval performance for `Qwen3-0.6B`and retention relative to its full-width Base model. Each cell reports the metric score on the first line and Base retention in parentheses on the second line. N and R denote NDCG@10 and Recall@10; MA denotes the Matryoshka Adapter. Bold green marks the best score among non-Base systems; bold black marks Base when it is best overall. Bold blue marks the highest retention among non-Base systems. Ties are highlighted at the displayed two-decimal precision

Language	Embedding system	$k=8$		$k=16$		$k=32$		$k=64$		$k=128$		$k=256$		$k=512$		$k=1024$	
		N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R
English	Base	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)	**60.71** (100.00%)	**65.59** (100.00%)
	MRL	5.67 (9.34%)	8.76 (13.36%)	19.26 (31.72%)	25.54 (38.93%)	40.42 (66.58%)	47.27 (72.07%)	50.03 (82.42%)	56.25 (85.75%)	55.67 (91.70%)	61.10 (93.15%)	58.47 (96.33%)	63.64 (97.03%)	**60.23** (**99.21%**)	**65.29** (**99.53%**)	**60.71** (**100.00%**)	**65.59** (**100.00%**)
	MA	3.79 (6.25%)	5.78 (8.81%)	16.48 (27.15%)	21.99 (33.53%)	35.88 (59.10%)	42.44 (64.71%)	48.25 (79.48%)	53.91 (82.18%)	55.14 (90.84%)	60.00 (91.48%)	58.04 (95.61%)	62.89 (95.88%)	59.81 (98.52%)	64.51 (98.35%)	60.57 (99.77%)	65.30 (99.55%)
	Expert	38.34 (63.16%)	45.31 (69.08%)	**47.27** (**77.86%**)	53.37 (81.37%)	**52.37** (**86.28%**)	**58.44** (**89.09%**)	55.34 (91.17%)	60.84 (92.75%)	57.29 (94.38%)	62.18 (94.80%)	58.44 (96.26%)	63.39 (96.65%)	59.07 (97.30%)	63.98 (97.53%)	59.29 (97.67%)	64.26 (97.97%)
	Japanese Expert	36.27 (59.75%)	44.32 (67.56%)	44.51 (73.33%)	51.87 (79.08%)	50.41 (83.03%)	56.20 (85.68%)	54.19 (89.27%)	60.26 (91.86%)	56.62 (93.28%)	62.06 (94.62%)	58.27 (95.99%)	63.03 (96.08%)	59.17 (97.46%)	64.27 (97.97%)	59.47 (97.97%)	64.43 (98.22%)
	Italian Expert	**38.37** (**63.20%**)	**45.35** (**69.13%**)	47.06 (77.52%)	**53.47** (**81.52%**)	51.12 (84.20%)	57.16 (87.14%)	54.55 (89.85%)	59.93 (91.36%)	57.05 (93.98%)	61.95 (94.45%)	58.42 (96.23%)	63.47 (96.76%)	59.30 (97.69%)	63.98 (97.54%)	59.43 (97.90%)	64.13 (97.77%)
	Hindi Expert	35.36 (58.24%)	42.88 (65.37%)	44.27 (72.93%)	51.37 (78.31%)	49.97 (82.32%)	56.39 (85.97%)	53.75 (88.54%)	59.56 (90.80%)	56.71 (93.42%)	62.03 (94.56%)	58.69 (96.69%)	**63.70** (**97.11%**)	59.26 (97.62%)	64.12 (97.75%)	59.40 (97.86%)	64.16 (97.82%)
	MORSE	36.16 (59.57%)	42.48 (64.75%)	46.18 (76.07%)	52.30 (79.73%)	52.36 (86.25%)	57.95 (88.35%)	**55.73** (**91.80%**)	**61.34** (**93.51%**)	**57.53** (**94.77%**)	**62.68** (**95.55%**)	**58.74** (**96.77%**)	63.51 (96.83%)	59.72 (98.38%)	64.47 (98.28%)	60.01 (98.85%)	64.76 (98.72%)
Japanese	Base	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	**72.07** (100.00%)	**74.10** (100.00%)	72.07 (100.00%)	74.10 (100.00%)
	MRL	10.15 (14.08%)	15.12 (20.40%)	27.98 (38.83%)	35.01 (47.25%)	49.63 (68.86%)	56.50 (76.25%)	60.59 (84.07%)	65.38 (88.24%)	66.26 (91.94%)	68.68 (92.68%)	68.98 (95.72%)	70.80 (95.55%)	70.40 (97.68%)	72.72 (98.14%)	72.07 (100.00%)	74.10 (100.00%)
	MA	7.04 (9.77%)	11.03 (14.89%)	22.70 (31.50%)	28.87 (38.96%)	44.21 (61.35%)	51.91 (70.05%)	58.08 (80.59%)	63.65 (85.90%)	65.81 (91.32%)	68.17 (92.00%)	68.66 (95.27%)	71.33 (96.26%)	70.71 (98.11%)	**74.01** (**99.88%**)	72.00 (99.90%)	**74.55** (**100.60%**)
	Expert	49.80 (69.10%)	56.61 (76.40%)	57.26 (79.46%)	62.30 (84.08%)	63.30 (87.83%)	67.90 (91.64%)	67.24 (93.30%)	**71.25** (**96.15%**)	69.39 (96.29%)	70.41 (95.02%)	69.24 (96.07%)	71.49 (96.48%)	70.00 (97.13%)	71.60 (96.62%)	69.76 (96.80%)	71.13 (95.99%)
	English Expert	49.12 (68.15%)	54.43 (73.45%)	**58.85** (**81.65%**)	62.85 (84.82%)	63.90 (88.66%)	66.39 (89.60%)	66.01 (91.59%)	70.02 (94.49%)	67.16 (93.20%)	69.74 (94.11%)	69.13 (95.92%)	71.18 (96.05%)	70.75 (98.17%)	72.19 (97.42%)	70.32 (97.58%)	72.99 (98.50%)
	Italian Expert	**50.92** (**70.66%**)	56.75 (76.59%)	58.19 (80.74%)	**64.56** (**87.13%**)	63.59 (88.24%)	68.66 (92.66%)	66.58 (92.38%)	70.25 (94.80%)	67.51 (93.68%)	70.27 (94.83%)	68.93 (95.65%)	71.33 (96.27%)	70.08 (97.24%)	71.36 (96.31%)	70.73 (98.14%)	72.13 (97.34%)
	Hindi Expert	50.46 (70.02%)	**56.86** (**76.74%**)	57.20 (79.37%)	62.32 (84.10%)	62.13 (86.22%)	65.15 (87.93%)	64.84 (89.98%)	67.76 (91.45%)	67.66 (93.88%)	70.05 (94.54%)	69.26 (96.11%)	71.40 (96.35%)	69.81 (96.86%)	72.40 (97.70%)	70.67 (98.05%)	72.53 (97.88%)
	MORSE	46.92 (65.11%)	53.34 (71.99%)	58.75 (81.52%)	63.46 (85.64%)	**65.25** (**90.54%**)	**69.95** (**94.40%**)	**67.30** (**93.38%**)	70.15 (94.67%)	**69.99** (**97.12%**)	**73.26** (**98.86%**)	**71.86** (**99.71%**)	**73.85** (**99.67%**)	**71.77** (**99.58%**)	73.58 (99.30%)	**72.10** (**100.05%**)	74.06 (99.95%)
Italian	Base	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	**73.93** (100.00%)	**73.79** (100.00%)	73.93 (100.00%)	73.79 (100.00%)	73.93 (100.00%)
	MRL	10.95 (14.84%)	16.61 (22.47%)	30.07 (40.75%)	39.40 (53.29%)	54.96 (74.48%)	60.11 (81.30%)	63.88 (86.57%)	68.18 (92.23%)	68.97 (93.46%)	71.84 (97.17%)	71.50 (96.89%)	73.15 (98.95%)	73.03 (98.96%)	74.04 (100.15%)	73.79 (**100.00%**)	73.93 (100.00%)
	MA	8.72 (11.82%)	14.77 (19.98%)	27.46 (37.21%)	34.58 (46.77%)	50.21 (68.04%)	56.24 (76.07%)	63.15 (85.58%)	66.44 (89.87%)	68.80 (93.23%)	69.93 (94.59%)	70.67 (95.77%)	71.78 (97.10%)	72.93 (98.83%)	73.38 (99.26%)	**73.80** (**100.00%**)	**74.69** (**101.03%**)
	Expert	**55.71** (**75.49%**)	**61.74** (**83.52%**)	63.18 (85.62%)	67.24 (90.95%)	67.65 (91.68%)	**70.82** (**95.79%**)	69.61 (94.33%)	**72.43** (**97.97%**)	71.03 (96.26%)	72.25 (97.72%)	71.67 (97.12%)	73.16 (98.96%)	73.00 (98.93%)	73.61 (99.57%)	73.31 (99.35%)	73.87 (99.92%)
	English Expert	52.76 (71.50%)	58.51 (79.14%)	**63.76** (**86.41%**)	**67.25** (**90.96%**)	66.62 (90.28%)	69.63 (94.19%)	69.86 (94.67%)	71.51 (96.73%)	**71.71** (**97.18%**)	72.32 (97.83%)	72.09 (97.70%)	73.26 (99.09%)	72.96 (98.87%)	73.23 (99.05%)	72.84 (98.71%)	73.36 (99.23%)
	Japanese Expert	50.91 (68.99%)	58.90 (79.67%)	58.95 (79.89%)	64.58 (87.35%)	64.70 (87.67%)	68.24 (92.31%)	67.98 (92.12%)	69.35 (93.81%)	69.95 (94.80%)	71.83 (97.16%)	71.95 (97.50%)	73.36 (99.23%)	72.88 (98.76%)	72.78 (98.45%)	72.99 (98.91%)	73.93 (100.00%)
	Hindi Expert	49.40 (66.95%)	56.00 (75.74%)	59.34 (80.42%)	64.60 (87.38%)	65.79 (89.15%)	69.76 (94.36%)	68.19 (92.41%)	70.73 (95.67%)	70.39 (95.38%)	72.00 (97.39%)	72.19 (97.82%)	**73.40** (**99.29%**)	71.99 (97.55%)	73.16 (98.96%)	72.84 (98.71%)	73.59 (99.54%)
	MORSE	52.37 (70.97%)	58.82 (79.56%)	63.02 (85.39%)	66.60 (90.09%)	**67.93** (**92.06%**)	70.40 (95.22%)	**70.24** (**95.18%**)	71.44 (96.64%)	71.64 (97.08%)	**72.58** (**98.18%**)	**72.46** (**98.19%**)	73.19 (99.00%)	**73.20** (**99.20%**)	**74.14** (**100.28%**)	73.38 (99.44%)	73.64 (99.60%)
Hindi	Base	**60.46** (100.00%)	**72.19** (100.00%)	**60.46** (100.00%)	**72.19** (100.00%)	**60.46** (100.00%)	**72.19** (100.00%)	**60.46** (100.00%)	**72.19** (100.00%)	**60.46** (100.00%)	**72.19** (100.00%)	**60.46** (100.00%)	**72.19** (100.00%)	60.46 (100.00%)	72.19 (100.00%)	60.46 (100.00%)	72.19 (100.00%)
	MRL	11.73 (19.41%)	20.51 (28.41%)	24.85 (41.10%)	36.72 (50.87%)	41.40 (68.48%)	54.59 (75.62%)	48.94 (80.94%)	61.65 (85.40%)	54.14 (89.55%)	66.25 (91.77%)	56.79 (93.94%)	68.69 (95.16%)	59.16 (97.85%)	70.95 (98.29%)	60.46 (100.00%)	72.19 (100.00%)
	MA	8.17 (13.51%)	14.53 (20.12%)	20.58 (34.05%)	30.97 (42.90%)	37.52 (62.05%)	49.63 (68.75%)	47.78 (79.03%)	59.22 (82.04%)	53.85 (89.06%)	65.60 (90.87%)	56.71 (93.80%)	68.43 (94.79%)	58.89 (97.41%)	70.67 (97.90%)	60.20 (99.57%)	71.97 (99.70%)
	Expert	**44.55** (**73.69%**)	**58.42** (**80.93%**)	**50.28** (**83.16%**)	**63.10** (**87.41%**)	53.76 (88.92%)	**66.60** (**92.27%**)	55.91 (92.47%)	67.98 (94.17%)	57.08 (94.41%)	69.42 (96.16%)	58.29 (96.41%)	70.20 (97.25%)	59.06 (97.68%)	71.11 (98.50%)	59.19 (97.90%)	71.10 (98.49%)
	English Expert	41.86 (69.24%)	55.50 (76.88%)	49.16 (81.31%)	62.69 (86.85%)	52.69 (87.16%)	65.54 (90.79%)	55.09 (91.12%)	67.73 (93.82%)	56.93 (94.17%)	69.09 (95.72%)	58.37 (96.54%)	70.29 (97.37%)	59.28 (97.79%)	71.30 (98.20%)	59.25 (98.00%)	70.81 (98.10%)
	Japanese Expert	42.67 (70.58%)	56.42 (78.16%)	48.85 (80.80%)	62.04 (85.94%)	52.60 (87.00%)	65.52 (90.76%)	55.27 (91.42%)	67.76 (93.87%)	56.94 (94.18%)	68.94 (95.50%)	58.52 (96.79%)	70.46 (97.60%)	59.28 (98.05%)	71.30 (98.78%)	59.41 (98.27%)	71.30 (98.77%)
	Italian Expert	42.80 (70.79%)	56.40 (78.13%)	49.52 (81.91%)	62.12 (86.06%)	52.70 (87.17%)	65.57 (90.84%)	55.13 (91.18%)	67.49 (93.49%)	56.81 (93.96%)	69.04 (95.64%)	58.24 (96.33%)	70.27 (97.35%)	58.91 (97.44%)	70.77 (98.04%)	59.20 (97.93%)	71.07 (98.46%)
	MORSE	42.47 (70.24%)	55.55 (76.95%)	49.90 (82.54%)	62.77 (86.95%)	**54.15** (**89.57%**)	**66.60** (92.26%)	**56.83** (**94.00%**)	**68.99** (**95.57%**)	**58.26** (**96.36%**)	**70.14** (**97.16%**)	**59.56** (**98.51%**)	**71.25** (**98.70%**)	**60.56** (**100.17%**)	**72.26** (**100.16%**)	**60.89** (**100.72%**)	**72.49** (**100.42%**)

Table F.4: MMTEB retrieval performance for `Qwen3-4B` and retention relative to its full-width Base model. Each cell reports the metric score on the first line and Base retention in parentheses on the second line. N and R denote NDCG@10 and Recall@10; MA denotes the Matryoshka Adapter. Bold green marks the best score among non-Base systems; bold black marks Base when it is best overall. Bold blue marks the highest retention among non-Base systems. Ties are highlighted at the displayed two-decimal precision

Language	Embedding system	$k = 8$		$k = 16$		$k = 32$		$k = 64$		$k = 128$		$k = 256$		$k = 512$		$k = 1024$		$k = 2560$	
		N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R
English	Base	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)	66.32 (100.00%)	**70.56** (100.00%)	**66.32** (100.00%)	**70.56** (100.00%)
	MRL	4.90 (7.39%)	7.72 (10.94%)	19.85 (29.93%)	25.71 (36.44%)	41.84 (63.09%)	48.97 (69.40%)	52.20 (78.71%)	58.27 (82.58%)	58.22 (87.78%)	63.63 (90.17%)	62.19 (93.77%)	67.44 (95.57%)	64.91 (97.87%)	**69.14** (**97.98%**)	**66.42** (**100.14%**)	**70.21** (**99.50%**)	**66.32** (**100.00%**)	**70.56** (**100.00%**)
	MA	3.36 (5.07%)	5.52 (7.82%)	15.06 (22.71%)	20.83 (29.52%)	36.51 (55.05%)	43.74 (61.99%)	49.62 (74.82%)	55.33 (78.41%)	57.14 (86.16%)	62.58 (88.69%)	60.31 (90.94%)	65.60 (92.97%)	62.96 (94.94%)	67.57 (95.75%)	64.86 (97.79%)	69.42 (98.39%)	65.61 (98.93%)	69.72 (98.80%)
	CSRv2	44.65 (67.32%)	**53.70** (**76.10%**)	51.04 (76.96%)	59.61 (84.48%)	56.15 (84.66%)	**64.38** (**91.24%**)	58.76 (88.60%)	**66.23** (**93.86%**)	60.29 (90.91%)	67.11 (95.11%)	61.63 (92.93%)	68.33 (96.83%)	62.46 (94.18%)	68.91 (97.65%)	–	–	62.82 (94.72%)	69.18 (98.04%)
	CSRv2-linear	44.75 (67.47%)	53.50 (75.82%)	51.12 (77.08%)	**59.78** (**84.72%**)	56.11 (84.59%)	64.03 (90.75%)	58.73 (88.55%)	66.06 (93.62%)	60.49 (91.20%)	67.47 (95.62%)	61.90 (93.33%)	**68.42** (**96.97%**)	62.63 (94.43%)	68.80 (97.50%)	–	–	62.75 (94.62%)	69.02 (97.81%)
	Expert	**44.86** (**67.63%**)	51.20 (72.55%)	**52.41** (**79.02%**)	58.59 (83.03%)	**58.18** (**87.72%**)	63.32 (89.74%)	60.84 (91.73%)	65.42 (92.71%)	62.52 (94.26%)	66.87 (94.76%)	63.70 (96.05%)	68.25 (96.72%)	64.66 (97.50%)	68.93 (97.69%)	65.13 (98.21%)	69.80 (98.92%)	65.39 (98.60%)	69.86 (99.01%)
	Japanese Expert	40.41 (60.92%)	47.60 (67.46%)	51.38 (77.47%)	57.13 (80.96%)	56.93 (85.84%)	61.94 (87.78%)	60.22 (90.80%)	64.70 (91.69%)	62.15 (93.71%)	66.66 (94.46%)	63.38 (95.57%)	68.06 (96.45%)	64.15 (96.73%)	68.75 (97.43%)	64.64 (97.46%)	69.17 (98.03%)	65.24 (98.37%)	69.47 (98.45%)
	Italian Expert	41.08 (61.94%)	48.05 (68.09%)	52.16 (78.64%)	57.97 (82.15%)	58.01 (87.46%)	63.38 (89.82%)	60.46 (91.17%)	65.27 (92.49%)	62.43 (94.13%)	66.93 (94.85%)	63.27 (95.39%)	68.08 (96.48%)	64.18 (96.77%)	68.64 (97.28%)	65.04 (98.07%)	69.67 (98.74%)	65.21 (98.32%)	69.59 (98.62%)
	Hindi Expert	42.20 (63.63%)	48.64 (68.93%)	51.79 (78.09%)	57.73 (81.81%)	56.61 (85.35%)	61.83 (87.63%)	59.31 (89.43%)	64.44 (91.32%)	61.60 (92.88%)	66.83 (94.71%)	63.29 (95.43%)	68.00 (96.37%)	64.13 (96.69%)	68.65 (97.28%)	64.76 (97.64%)	69.58 (98.61%)	65.06 (98.10%)	69.74 (98.83%)
	MORSE	42.11 (63.49%)	48.40 (68.59%)	52.40 (79.00%)	57.83 (81.96%)	58.14 (87.66%)	62.96 (89.23%)	**61.03** (**92.02%**)	65.66 (93.05%)	**63.22** (**95.32%**)	**67.52** (**95.69%**)	**64.17** (**96.75%**)	68.20 (96.65%)	**65.03** (**98.04%**)	69.09 (97.92%)	65.52 (98.78%)	69.40 (98.35%)	65.47 (98.72%)	69.39 (98.34%)
Japanese	Base	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)	**78.12** (100.00%)	**78.83** (100.00%)
	MRL	9.41 (12.04%)	14.01 (17.77%)	29.71 (38.02%)	37.40 (47.44%)	52.07 (66.66%)	57.62 (73.09%)	60.96 (78.04%)	64.01 (81.20%)	68.10 (87.17%)	71.50 (90.70%)	72.49 (92.79%)	73.50 (93.23%)	76.62 (98.07%)	**77.87** (**98.78%**)	77.35 (99.02%)	78.09 (99.06%)	**78.12** (**100.00%**)	**78.83** (**100.00%**)
	MA	6.55 (8.38%)	11.47 (14.55%)	24.13 (30.89%)	31.33 (39.75%)	48.25 (61.76%)	53.77 (68.21%)	62.98 (80.62%)	67.43 (85.53%)	70.81 (90.64%)	73.49 (93.22%)	73.43 (93.99%)	75.20 (95.39%)	76.01 (97.29%)	76.57 (97.13%)	76.21 (97.55%)	77.39 (98.17%)	77.38 (99.05%)	77.64 (98.49%)
	Expert	60.44 (77.37%)	66.09 (83.83%)	**68.55** (**87.74%**)	70.35 (89.24%)	71.64 (91.70%)	73.49 (93.22%)	72.20 (92.42%)	74.24 (94.17%)	75.24 (96.31%)	**76.43** (**96.95%**)	75.34 (96.44%)	76.60 (97.16%)	75.90 (97.15%)	77.10 (97.80%)	76.14 (97.46%)	76.85 (97.48%)	76.66 (98.13%)	77.78 (98.67%)
	English Expert	**61.15** (**78.27%**)	**66.81** (**84.75%**)	68.24 (87.35%)	**70.87** (**89.90%**)	72.52 (92.83%)	**74.20** (**94.13%**)	73.09 (93.56%)	74.98 (95.11%)	**75.36** (**96.46%**)	76.42 (96.94%)	74.78 (95.71%)	75.34 (95.57%)	76.28 (97.64%)	76.01 (96.42%)	75.52 (96.67%)	76.73 (97.34%)	75.94 (97.21%)	76.35 (96.85%)
	Italian Expert	56.94 (72.88%)	63.43 (80.46%)	67.44 (86.32%)	70.42 (89.33%)	71.79 (91.90%)	73.55 (93.29%)	73.30 (93.82%)	**75.91** (**96.29%**)	74.47 (95.32%)	75.95 (96.34%)	75.96 (97.23%)	76.37 (96.88%)	75.52 (96.66%)	76.58 (97.14%)	76.12 (97.43%)	77.38 (98.15%)	76.24 (97.59%)	77.23 (97.97%)
	Hindi Expert	56.11 (71.82%)	62.94 (79.83%)	66.84 (85.55%)	70.51 (89.44%)	71.20 (91.14%)	72.11 (91.47%)	71.71 (91.78%)	72.95 (92.54%)	73.30 (93.83%)	74.58 (94.60%)	75.67 (96.86%)	76.76 (97.38%)	75.95 (97.21%)	76.80 (97.42%)	76.16 (97.49%)	75.96 (96.35%)	76.28 (97.65%)	76.88 (97.52%)
	MORSE	57.83 (74.02%)	63.21 (80.18%)	67.01 (85.78%)	69.74 (88.46%)	**72.93** (**93.36%**)	73.05 (92.67%)	**73.85** (**94.53%**)	74.41 (94.39%)	74.59 (95.47%)	75.53 (95.80%)	**76.19** (**97.52%**)	**77.23** (**97.97%**)	**76.93** (**98.47%**)	76.96 (97.62%)	**77.62** (**99.36%**)	**78.18** (**99.16%**)	77.82 (99.60%)	78.22 (99.22%)
Italian	Base	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)	79.48 (100.00%)	79.70 (100.00%)	**79.48** (100.00%)	**79.70** (100.00%)
	MRL	11.36 (14.30%)	18.28 (22.94%)	34.19 (43.02%)	41.89 (52.55%)	57.13 (71.87%)	62.97 (79.00%)	65.79 (82.77%)	69.71 (87.46%)	72.28 (90.94%)	74.55 (93.53%)	75.55 (95.06%)	77.51 (97.24%)	78.75 (99.09%)	**78.89** (**98.98%**)	**79.82** (**100.43%**)	**79.71** (**100.01%**)	**79.48** (**100.00%**)	**79.70** (**100.00%**)
	MA	8.56 (10.77%)	12.74 (15.99%)	27.62 (34.75%)	34.64 (43.47%)	53.25 (66.99%)	60.32 (75.68%)	65.66 (82.61%)	69.90 (87.70%)	71.12 (89.48%)	74.33 (93.25%)	73.35 (92.29%)	75.56 (94.80%)	75.59 (95.11%)	77.25 (96.91%)	76.78 (96.60%)	78.27 (98.20%)	77.07 (96.96%)	78.17 (98.07%)
	Expert	60.78 (76.47%)	**65.11** (**81.68%**)	67.51 (84.94%)	70.93 (88.99%)	73.39 (92.34%)	74.46 (93.43%)	75.19 (94.60%)	**76.80** (**96.35%**)	75.10 (94.48%)	77.29 (96.97%)	77.12 (97.04%)	78.28 (98.21%)	77.51 (97.52%)	78.30 (98.24%)	78.19 (98.38%)	79.31 (99.51%)	77.77 (97.85%)	78.67 (98.70%)
	English Expert	**60.81** (**76.50%**)	63.77 (80.00%)	**68.96** (**86.77%**)	**72.61** (**91.10%**)	73.28 (92.20%)	**75.03** (**94.13%**)	74.90 (94.23%)	75.18 (94.33%)	76.52 (96.28%)	**77.61** (**97.37%**)	76.82 (96.65%)	76.74 (96.28%)	76.62 (96.40%)	76.94 (96.53%)	77.81 (97.90%)	78.29 (98.22%)	77.97 (98.09%)	78.60 (98.61%)
	Japanese Expert	54.07 (68.03%)	62.66 (78.62%)	66.78 (84.02%)	70.14 (88.00%)	70.97 (89.30%)	72.60 (91.08%)	73.29 (92.21%)	75.03 (94.14%)	76.03 (95.65%)	77.00 (96.61%)	76.58 (96.35%)	77.03 (96.64%)	77.12 (97.03%)	77.82 (97.63%)	77.25 (97.20%)	78.09 (97.97%)	77.57 (97.59%)	78.71 (98.75%)
	Hindi Expert	57.84 (72.77%)	63.11 (79.17%)	68.11 (85.69%)	71.92 (90.23%)	70.79 (89.07%)	74.28 (93.20%)	74.37 (93.57%)	75.35 (94.53%)	75.36 (94.82%)	76.27 (95.69%)	76.84 (96.68%)	76.97 (96.57%)	77.94 (98.05%)	78.11 (97.99%)	77.96 (98.09%)	78.36 (98.32%)	77.77 (97.85%)	79.07 (99.21%)
	MORSE	59.66 (75.06%)	62.68 (78.63%)	67.90 (85.44%)	69.69 (87.43%)	**73.67** (**92.68%**)	74.09 (92.96%)	**76.23** (**95.91%**)	76.15 (95.54%)	**76.70** (**96.50%**)	77.05 (96.67%)	**77.74** (**97.81%**)	**78.38** (**98.34%**)	**78.77** (**99.10%**)	78.81 (98.88%)	79.49 (100.02%)	79.03 (99.15%)	79.02 (99.41%)	78.45 (98.43%)
Hindi	Base	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)	**74.38** (100.00%)	**87.00** (100.00%)
	MRL	12.91 (17.35%)	22.95 (26.38%)	32.36 (43.51%)	47.51 (54.61%)	53.74 (72.26%)	69.25 (79.60%)	60.70 (81.62%)	75.26 (86.50%)	65.90 (88.60%)	79.62 (91.52%)	69.13 (92.95%)	82.39 (94.70%)	71.72 (96.43%)	84.48 (97.11%)	73.52 (98.84%)	86.18 (99.07%)	**74.38** (**100.00%**)	**87.00** (**100.00%**)
	MA	7.41 (9.97%)	13.68 (15.72%)	21.60 (29.04%)	32.88 (37.80%)	42.29 (56.85%)	56.56 (65.02%)	56.44 (75.88%)	70.21 (80.70%)	63.65 (85.58%)	77.25 (88.80%)	66.97 (90.05%)	80.36 (92.37%)	69.98 (94.08%)	83.32 (95.78%)	71.76 (96.48%)	84.90 (97.59%)	73.01 (98.16%)	86.00 (98.86%)
	Expert	55.96 (75.24%)	71.53 (82.23%)	**62.86** (**84.52%**)	77.51 (89.09%)	66.51 (89.42%)	80.51 (92.54%)	68.69 (92.35%)	82.41 (94.73%)	70.02 (94.14%)	83.41 (95.87%)	71.40 (96.00%)	84.69 (97.34%)	72.16 (97.01%)	85.26 (98.00%)	72.82 (97.90%)	85.52 (98.30%)	73.02 (98.18%)	85.92 (98.76%)
	English Expert	**56.27** (**75.65%**)	**71.56** (**82.26%**)	62.77 (84.39%)	**77.57** (**89.17%**)	66.26 (89.09%)	80.60 (92.65%)	68.08 (91.53%)	81.94 (94.18%)	69.78 (93.82%)	83.40 (95.87%)	70.78 (95.16%)	84.00 (96.56%)	71.47 (96.08%)	84.69 (97.34%)	72.47 (97.43%)	85.27 (98.02%)	72.77 (97.84%)	85.75 (98.57%)
	Japanese Expert	54.13 (72.77%)	69.30 (79.66%)	61.96 (83.30%)	76.89 (88.38%)	65.50 (88.07%)	79.86 (91.80%)	68.03 (91.46%)	82.04 (94.30%)	69.66 (93.66%)	83.18 (95.61%)	70.73 (95.09%)	84.05 (96.61%)	71.76 (96.49%)	84.81 (97.49%)	72.39 (97.33%)	85.36 (98.12%)	72.69 (97.73%)	85.45 (98.22%)
	Hindi Expert	54.41 (73.15%)	70.12 (80.60%)	61.94 (83.28%)	77.19 (88.73%)	65.98 (88.71%)	80.04 (92.00%)	67.87 (91.25%)	81.59 (93.78%)	69.70 (93.71%)	82.94 (95.34%)	70.83 (95.23%)	84.09 (96.66%)	71.59 (96.25%)	84.94 (97.63%)	72.39 (97.32%)	85.36 (98.12%)	72.61 (97.62%)	85.56 (98.35%)
	MORSE	54.61 (73.42%)	69.65 (80.06%)	62.74 (84.35%)	76.91 (88.41%)	**67.01** (**90.10%**)	**80.95** (**93.05%**)	**69.73** (**93.75%**)	**83.07** (**95.48%**)	**71.53** (**96.17%**)	**84.56** (**97.19%**)	**72.69** (**97.73%**)	**85.31** (**98.06%**)	**73.54** (**98.87%**)	**85.97** (**98.82%**)	**74.06** (**99.57%**)	**86.48** (**99.41%**)	74.32 (99.93%)	86.71 (99.67%)

Table F.5: MMTEB retrieval performance for `bge-m3` and retention relative to its full-width Base model. Each cell reports the metric score on the first line and Base retention in parentheses on the second line. N and R denote NDCG@10 and Recall@10; MA denotes the Matryoshka Adapter. Bold green marks the best score among non-Base systems; bold black marks Base when it is best overall. Bold blue marks the highest retention among non-Base systems. Ties are highlighted at the displayed two-decimal precision

Language	Embedding system	$k = 8$		$k = 16$		$k = 32$		$k = 64$		$k = 128$		$k = 256$		$k = 512$		$k = 1024$	
		N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R
English	Base	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)	**56.11** (100.00%)	**60.94** (100.00%)
	First-k	2.23 (3.97%)	3.55 (5.82%)	7.03 (12.53%)	11.14 (18.29%)	17.89 (31.89%)	24.16 (39.64%)	30.96 (55.18%)	37.47 (61.49%)	42.95 (76.54%)	48.36 (79.36%)	50.79 (90.53%)	55.89 (91.72%)	54.43 (97.02%)	59.51 (97.65%)	**56.11** (**100.00%**)	**60.94** (**100.00%**)
	Top-k	3.63 (6.47%)	5.58 (9.16%)	9.69 (17.26%)	12.43 (20.40%)	22.16 (39.50%)	28.37 (46.55%)	36.70 (65.42%)	42.96 (70.49%)	47.48 (84.62%)	52.90 (86.80%)	52.88 (94.25%)	58.25 (95.58%)	**55.54** (**98.99%**)	60.41 (99.13%)	**56.11** (**100.00%**)	**60.94** (**100.00%**)
	MA	1.61 (2.87%)	2.74 (4.49%)	6.65 (11.85%)	9.81 (16.10%)	19.42 (34.61%)	25.51 (41.86%)	34.38 (61.27%)	40.38 (66.26%)	43.87 (78.18%)	49.10 (80.57%)	50.19 (89.46%)	55.52 (91.10%)	53.77 (95.83%)	59.00 (96.81%)	55.75 (99.36%)	60.50 (99.28%)
	Expert	**29.20** (**52.05%**)	**35.33** (**57.97%**)	**39.93** (**71.16%**)	**46.82** (**76.82%**)	**45.64** (**81.34%**)	**52.17** (**85.61%**)	49.97 (89.06%)	**56.00** (**91.90%**)	51.94 (92.58%)	58.22 (95.53%)	53.50 (95.36%)	59.25 (97.22%)	54.67 (97.43%)	59.96 (98.39%)	55.10 (98.20%)	60.21 (98.79%)
	Japanese Expert	27.61 (49.21%)	33.99 (55.78%)	36.99 (65.94%)	43.83 (71.92%)	43.55 (77.61%)	49.83 (81.77%)	48.64 (86.70%)	55.27 (90.70%)	51.78 (92.28%)	58.22 (95.53%)	54.03 (96.31%)	59.33 (97.36%)	54.66 (97.42%)	59.96 (98.39%)	54.99 (98.00%)	59.99 (98.44%)
	Italian Expert	29.19 (52.03%)	34.96 (57.36%)	38.58 (68.76%)	44.48 (72.98%)	45.05 (80.29%)	51.30 (84.18%)	48.67 (86.74%)	55.58 (91.19%)	51.91 (92.52%)	58.04 (95.24%)	53.94 (96.14%)	59.46 (97.56%)	54.67 (97.44%)	60.16 (98.72%)	55.04 (98.10%)	60.39 (99.09%)
	Hindi Expert	27.19 (48.45%)	33.27 (54.60%)	36.57 (65.18%)	42.34 (69.48%)	43.58 (77.68%)	49.23 (80.79%)	48.13 (85.79%)	53.56 (87.89%)	51.55 (91.88%)	57.67 (94.62%)	53.75 (95.81%)	59.34 (97.37%)	55.03 (98.08%)	60.16 (98.72%)	55.27 (98.50%)	60.45 (99.19%)
	MORSE	25.82 (46.01%)	31.60 (51.85%)	36.73 (65.46%)	43.06 (70.66%)	44.84 (79.93%)	50.61 (83.04%)	**50.21** (**89.48%**)	55.79 (91.54%)	**53.69** (**95.69%**)	**58.68** (**96.29%**)	**55.08** (**98.17%**)	**60.23** (**98.83%**)	55.47 (98.87%)	**60.58** (**99.40%**)	55.53 (98.97%)	60.67 (99.56%)
Japanese	Base	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)	**75.17** (100.00%)	**77.31** (100.00%)
	First-k	4.22 (5.61%)	7.90 (10.21%)	12.57 (16.72%)	18.94 (24.50%)	26.03 (34.63%)	34.02 (44.00%)	44.77 (59.56%)	51.30 (66.36%)	61.55 (81.88%)	66.91 (86.55%)	69.82 (92.88%)	73.33 (94.86%)	73.50 (97.77%)	75.98 (98.29%)	**75.17** (**100.00%**)	**77.31** (**100.00%**)
	Top-k	8.60 (11.44%)	11.29 (14.60%)	17.20 (22.88%)	20.68 (26.75%)	33.90 (45.10%)	38.96 (50.40%)	52.87 (70.33%)	59.94 (77.53%)	66.58 (88.57%)	70.73 (91.50%)	71.96 (95.73%)	75.60 (97.79%)	**74.67** (**99.34%**)	**76.92** (**99.50%**)	**75.17** (**100.00%**)	**77.31** (**100.00%**)
	MA	4.03 (5.36%)	6.86 (8.88%)	11.90 (15.84%)	17.51 (22.65%)	31.85 (42.37%)	39.83 (51.52%)	50.10 (66.65%)	58.08 (75.14%)	62.96 (83.76%)	68.76 (88.94%)	69.98 (93.10%)	73.47 (95.04%)	73.61 (97.92%)	76.34 (98.76%)	74.78 (99.47%)	76.67 (99.17%)
	Expert	44.98 (59.83%)	**52.33** (**67.69%**)	**57.22** (**76.12%**)	**63.95** (**82.73%**)	**63.05** (**83.87%**)	**69.79** (**90.28%**)	67.40 (89.65%)	72.04 (93.19%)	69.01 (91.80%)	73.32 (94.84%)	71.69 (95.37%)	75.04 (97.07%)	71.80 (95.51%)	74.78 (96.73%)	72.10 (95.91%)	75.33 (97.44%)
	English Expert	43.13 (57.37%)	50.99 (65.96%)	55.31 (73.57%)	61.39 (79.42%)	61.81 (82.23%)	68.91 (89.14%)	66.29 (88.19%)	71.54 (92.54%)	70.07 (93.21%)	74.18 (95.96%)	71.55 (95.18%)	74.57 (96.46%)	71.50 (95.12%)	75.45 (97.60%)	71.79 (95.51%)	75.20 (97.27%)
	Italian Expert	**44.99** (**59.85%**)	51.24 (66.28%)	53.32 (70.93%)	60.66 (78.46%)	61.79 (82.19%)	68.01 (87.98%)	64.77 (86.16%)	71.75 (92.81%)	69.79 (92.84%)	73.19 (94.68%)	70.40 (93.65%)	74.22 (96.01%)	72.32 (96.21%)	75.52 (97.69%)	72.12 (95.93%)	75.42 (97.56%)
	Hindi Expert	43.28 (57.58%)	50.08 (64.78%)	54.41 (72.38%)	61.54 (79.61%)	60.81 (80.89%)	67.30 (87.05%)	66.84 (88.92%)	72.70 (94.04%)	69.11 (91.93%)	72.58 (93.89%)	72.32 (96.20%)	74.79 (96.74%)	72.97 (97.08%)	75.81 (98.06%)	73.10 (97.24%)	76.21 (98.58%)
	MORSE	40.11 (53.36%)	46.55 (60.21%)	53.23 (70.81%)	61.79 (79.93%)	62.67 (83.37%)	68.62 (88.76%)	**68.85** (**91.59%**)	**73.74** (**95.39%**)	**72.08** (**95.89%**)	**75.93** (**98.22%**)	**73.03** (**97.15%**)	**76.08** (**98.41%**)	73.96 (98.39%)	76.73 (99.26%)	74.32 (98.87%)	76.68 (99.19%)
Italian	Base	**75.37** (100.00%)	**75.81** (100.00%)	**75.37** (100.00%)	**75.81** (100.00%)	**75.37** (100.00%)	**75.81** (100.00%)	**75.37** (100.00%)	**75.81** (100.00%)	**75.37** (100.00%)	**75.81** (100.00%)	**75.37** (100.00%)	**75.81** (100.00%)	75.37 (100.00%)	75.81 (100.00%)	75.37 (100.00%)	75.81 (100.00%)
	First-k	4.73 (6.28%)	7.48 (9.86%)	13.13 (17.43%)	18.53 (24.44%)	28.48 (37.79%)	37.76 (49.81%)	46.80 (62.09%)	55.70 (73.47%)	62.52 (82.95%)	66.74 (88.04%)	70.57 (93.64%)	72.85 (96.10%)	73.80 (97.92%)	75.30 (99.33%)	75.37 (100.00%)	75.81 (100.00%)
	Top-k	6.06 (8.05%)	9.30 (12.27%)	16.57 (21.98%)	22.92 (30.23%)	35.31 (46.85%)	42.89 (56.57%)	55.00 (72.97%)	60.30 (79.54%)	66.10 (87.70%)	71.11 (93.79%)	72.78 (96.56%)	74.89 (98.78%)	75.15 (99.71%)	75.50 (99.58%)	75.37 (100.00%)	75.81 (100.00%)
	MA	2.79 (3.71%)	4.72 (6.22%)	12.38 (16.43%)	18.44 (24.32%)	31.85 (42.25%)	38.61 (50.93%)	50.09 (66.46%)	58.80 (77.56%)	62.74 (83.24%)	66.85 (88.18%)	69.44 (92.13%)	71.99 (94.96%)	73.30 (97.25%)	74.79 (98.65%)	74.49 (98.84%)	76.09 (100.37%)
	Expert	**47.72** (**63.31%**)	**55.35** (**73.01%**)	**60.05** (**79.67%**)	**65.26** (**86.09%**)	65.23 (86.55%)	69.59 (91.79%)	69.00 (91.55%)	72.19 (95.22%)	71.70 (95.13%)	74.34 (98.06%)	72.91 (96.74%)	74.21 (97.89%)	74.16 (98.40%)	75.19 (99.19%)	74.32 (98.61%)	75.56 (99.68%)
	English Expert	46.10 (61.17%)	54.49 (71.88%)	58.50 (77.62%)	64.23 (84.72%)	65.40 (86.78%)	**70.60** (**93.13%**)	67.69 (89.81%)	71.52 (94.34%)	71.26 (94.54%)	73.34 (96.75%)	73.02 (96.88%)	74.26 (97.96%)	73.80 (97.92%)	74.61 (98.41%)	74.06 (98.26%)	74.63 (98.45%)
	Japanese Expert	45.60 (60.50%)	50.83 (67.05%)	57.16 (75.84%)	62.05 (81.85%)	63.70 (84.51%)	67.37 (88.87%)	68.26 (90.56%)	72.21 (95.26%)	71.58 (94.97%)	74.11 (97.76%)	73.63 (97.69%)	**75.10** (**99.06%**)	73.97 (98.15%)	75.07 (99.03%)	74.24 (98.51%)	75.35 (99.40%)
	Hindi Expert	42.94 (56.97%)	51.08 (67.38%)	55.97 (74.27%)	61.06 (80.55%)	63.01 (83.61%)	66.33 (87.50%)	68.14 (90.41%)	69.96 (92.28%)	71.73 (95.17%)	73.31 (96.70%)	73.31 (97.27%)	74.70 (98.54%)	74.26 (98.53%)	75.26 (99.27%)	74.47 (98.81%)	75.36 (99.41%)
	MORSE	42.37 (56.22%)	49.28 (65.00%)	57.52 (76.32%)	61.47 (81.09%)	**66.68** (**88.47%**)	70.27 (92.69%)	**71.62** (**95.02%**)	**73.24** (**96.61%**)	**73.88** (**98.02%**)	**74.59** (**98.39%**)	**74.67** (**99.08%**)	74.84 (98.71%)	**75.91** (**100.71%**)	**76.11** (**100.39%**)	**75.87** (**100.67%**)	**76.13** (**100.42%**)
Hindi	Base	**64.64** (100.00%)	**75.55** (100.00%)	**64.64** (100.00%)	**75.55** (100.00%)	**64.64** (100.00%)	**75.55** (100.00%)	**64.64** (100.00%)	**75.55** (100.00%)	**64.64** (100.00%)	**75.55** (100.00%)	**64.64** (100.00%)	**75.55** (100.00%)	64.55 (100.00%)	75.55 (100.00%)	**64.64** (100.00%)	75.55 (100.00%)
	First-k	7.56 (11.70%)	13.61 (18.02%)	15.48 (23.95%)	23.63 (31.28%)	26.54 (41.06%)	37.50 (49.64%)	40.11 (62.05%)	52.16 (69.04%)	51.17 (79.16%)	63.04 (83.45%)	58.77 (90.92%)	69.90 (92.52%)	62.35 (96.45%)	73.61 (97.43%)	**64.64** (**100.00%**)	75.55 (100.00%)
	Top-k	7.22 (11.17%)	11.83 (15.66%)	15.41 (23.85%)	22.28 (29.49%)	28.12 (43.51%)	38.55 (51.02%)	43.99 (68.06%)	55.53 (73.50%)	54.93 (84.98%)	66.48 (88.00%)	61.62 (95.33%)	72.65 (96.17%)	64.13 (99.21%)	75.02 (99.31%)	**64.64** (**100.00%**)	75.55 (100.00%)
	MA	6.36 (9.84%)	11.57 (15.32%)	16.01 (24.76%)	24.54 (32.48%)	29.73 (45.99%)	41.35 (54.75%)	43.35 (67.07%)	56.28 (74.49%)	52.41 (81.08%)	64.75 (85.71%)	58.52 (90.53%)	70.00 (92.66%)	61.89 (95.75%)	73.17 (96.86%)	64.23 (99.37%)	75.24 (99.59%)
	Expert	43.01 (66.53%)	**57.02** (**75.47%**)	**51.34** (**79.42%**)	**65.37** (**86.53%**)	**56.11** (**86.81%**)	**69.19** (**91.58%**)	59.41 (91.92%)	**71.86** (**95.12%**)	61.90 (95.76%)	**74.04** (**98.01%**)	63.46 (98.18%)	74.91 (99.15%)	64.39 (99.62%)	**75.68** (**100.17%**)	64.50 (99.79%)	**75.94** (**100.52%**)
	English Expert	**43.12** (**66.71%**)	**57.02** (**75.47%**)	50.98 (78.87%)	64.39 (85.23%)	56.08 (86.76%)	69.15 (91.53%)	59.35 (91.81%)	71.54 (94.69%)	61.73 (95.50%)	73.67 (97.52%)	63.33 (97.97%)	74.54 (98.66%)	64.14 (99.23%)	75.45 (99.87%)	64.29 (99.45%)	75.53 (99.98%)
	Japanese Expert	43.06 (66.62%)	56.93 (75.36%)	50.48 (78.09%)	64.14 (84.90%)	55.23 (85.45%)	68.24 (90.32%)	58.86 (91.06%)	71.19 (94.24%)	61.69 (95.43%)	73.63 (97.46%)	63.34 (98.00%)	75.01 (99.29%)	64.08 (99.14%)	75.39 (99.79%)	64.15 (99.25%)	75.47 (99.90%)
	Hindi Expert	43.09 (66.66%)	56.96 (75.40%)	50.92 (78.77%)	64.52 (85.41%)	55.94 (86.55%)	68.86 (91.14%)	59.29 (91.72%)	71.42 (94.54%)	61.74 (95.52%)	73.68 (97.52%)	63.46 (98.18%)	**75.02** (**99.30%**)	64.11 (99.18%)	75.52 (99.97%)	64.42 (99.67%)	75.86 (100.41%)
	MORSE	39.19 (60.63%)	52.55 (69.56%)	48.53 (75.08%)	61.47 (81.36%)	54.96 (85.02%)	67.06 (88.77%)	**59.58** (**92.18%**)	71.33 (94.42%)	**62.47** (**96.64%**)	73.56 (97.37%)	**63.89** (**98.85%**)	75.00 (99.27%)	**64.49** (**99.77%**)	75.41 (99.81%)	64.62 (99.97%)	75.33 (99.71%)