LEARNING RETRIEVAL MODELS WITH SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) provide a powerful mechanism for decomposing the dense representations produced by Large Language Models (LLMs) into interpretable latent features. We posit that SAEs constitute a natural foundation for Learned Sparse Retrieval (LSR), whose objective is to encode queries and documents into high-dimensional sparse representations optimized for efficient retrieval. In contrast to existing LSR approaches that project input sequences into the vocabulary space, SAE-based representations offer the potential to produce more semantically structured, expressive, and language-agnostic features. By leveraging recently released open-source SAEs, we show that their latent features can serve as effective indexing units for representing documents and queries for sparse retrieval. Our experiments demonstrate that SAE-based LSR models consistently outperform their vocabulary-based counterparts in multilingual and out-of-domain settings. Finally, we introduce SPLARE, a 7B-parameter multilingual retrieval model capable of producing generalizable sparse latent embeddings for a wide range of languages and domains, achieving top results on MMTEB's multilingual and English retrieval tasks. We also release a more efficient 2B-parameter variant, offering strong performance with a significantly lighter footprint.

1 INTRODUCTION

Embedding models have become a pivotal tool for search systems, enabling the better capture of semantic relationships between queries and documents across various domains and modalities. This trend has been further accelerated by the advent of Retrieval-Augmented Generation (RAG) Lewis et al. (2020) and agent-based systems, which impose even higher demands on retrieval performance and robustness. Recently, dense embedding models Reimers & Gurevych (2019); Karpukhin et al. (2020), which map inputs into single dense vectors, have demonstrated impressive performance on the (M)MTEB benchmark (Muennighoff et al., 2023; Enevoldsen et al., 2025). Specifically, embedding models relying on large (V)LLM backbones have become the de-facto approach for generalist multilingual Lee et al. (2025b); Zhang et al. (2025); Wang et al. (2024a); Lee et al. (2025a); Li et al. (2023b) or even multi-modal models Günther et al. (2025); Faysse et al. (2025); Xu et al. (2025a)—marking a shift away from encoder-only language models which have defined the state of the art for years (Izacard et al., 2022; Karpukhin et al., 2020; Xiong et al., 2020).

Learned Sparse Retrieval (LSR) methods (Formal et al., 2021; Mallia et al., 2021; Nguyen et al., 2023; Kong et al., 2023) have achieved state-of-the-art performance on widely used English-centric benchmarks (Thakur et al., 2021; Bajaj et al., 2018; Craswell et al., 2021) and have demonstrated strong generalization when compared to dense embedding models (Formal et al., 2022b; Lupart et al., 2023; Déjean et al., 2023). Beyond their efficiency, these approaches provide a level of interpretability that is particularly valuable in production systems. Models such as SPLADE (Formal et al., 2021; 2022a; Lassance et al., 2024) operationalize this idea by representing documents and queries as sparse, weighted bag-of-words over the vocabulary space of their backbone model. While originally developed for encoder-only architectures such as BERT (Devlin et al., 2019), recent work has explored adapting SPLADE to LLM backbones (Qiao et al., 2025; Doshi et al., 2024; Xu et al., 2025b; Zeng et al., 2025; Soares et al., 2023; Ma et al., 2025). However, these models remain limited to English-centric contexts and struggle to match state-of-the-art performance on more comprehensive benchmarks like MMTEB which place greater emphasis on generalization across novel domains and languages. Unlike dense retrieval, which models relevance within a continuous em-

bedding space, LSR methods are inherently constrained by the fixed vocabulary of their underlying backbone, which incurs issues such as tokenization redundancy Lei et al. (2025). This limitation also makes it significantly harder to handle multilingual or cross-lingual retrieval Nair et al. (2023; 2022); Lassance (2023)—and even more so when extending to multi-modal settings (Nguyen et al., 2024). We hypothesize that this is the primary reason why LSR models have recently fallen behind dense approaches[1].

In the context of LLMs, Sparse Autoencoders (SAEs) Makhzani & Frey (2013); Huben et al. (2024); Bricken et al. (2023) decompose dense token representations into sparse vectors of latent features. These features have been shown to exhibit desirable properties: they are largely mono-semantic (most features correspond to a single interpretable concept), multilingual (remaining largely language-agnostic), and even multimodal (generalizing across modalities in multimodal LLMs) (Bricken et al., 2023; Templeton et al., 2024; Lieberum et al., 2024; Huben et al., 2024; He et al., 2024; Deng et al., 2025). While SAEs have generated significant excitement for mechanistic interpretability, recent work has also highlighted their limitations, showing that they can struggle to transfer effectively to certain downstream tasks (Kantamneni et al., 2025; Smith et al., 2025).

In this work, we argue and empirically demonstrate that SAEs are a natural fit for LSR models: their learned latent features provide a semantically-grounded representation space for sparse retrieval which is particularly advantageous in domains or languages where vocabulary-based approaches may underperform. To this end, we propose a new LSR approach that represents queries and documents as sparse vectors over a latent vocabulary space, by replacing the standard language modeling (LM) head with pre-trained SAEs such as Llama Scope (He et al., 2024). More specifically, our contributions are as follows:

- We introduce SPLARE—for SParse LAtent REtrieval—a new LSR approach relying on pre-trained SAEs;
- We conduct a systematic investigation of the advantages of using a latent vocabulary—compared to the standard LLM vocabulary—across a comprehensive set of benchmarks spanning diverse tasks, domains, and languages;
- Finally, we introduce a new 7B multilingual latent sparse retriever that support 100+ languages and achieves competitive results on the MMTEB *retrieval* benchmark[2]. SPLARE is the first LSR model to rival state-of-the-art dense approaches on MMTEB. We additionally release a compact and efficient 2B counterpart.

2 Background

We first provide some background on sparse autoencoders as well as Learned Sparse Retrieval. SPLARE can be understood as synthesizing these two research directions into a unified framework.

2.1 Sparse Autoencoders

Given activations $x \in \mathbb{R}^d$ from a language model, a sparse autoencoder (SAE) is a single hidden layer model, comprising an encoder and a decoder:

$$z = f(\boldsymbol{W}_{\text{enc}} x + \boldsymbol{b}_{\text{enc}}), \quad \hat{x} = \boldsymbol{W}_{\text{dec}} z + \boldsymbol{b}_{\text{dec}} \tag{1}$$

where $z \in \mathbb{R}^{|\mathcal{W}|}$, with $|\mathcal{W}| >> d$ corresponding to the width of SAE, i.e., the number of features in the latent space. SAEs, as a class of autoencoders, are trained using a standard reconstruction objective $\mathcal{L} = \|\hat{x} - x\|^2$. Sparsity in the decomposition is induced through suitable activation functions f such as ReLU Bricken et al. (2023), Top-K Makhzani & Frey (2013); Gao et al. (2025) or JumpReLU Rajamanoharan et al. (2024), and regularization penalties such as ℓ_1. Several works have demonstrated that SAEs can recover highly monosemantic features, many of which are language-agnostic—responding consistently to the same concepts across languages—and, in some cases, even multimodal (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024; Lieberum et al., 2024;

[1] For instance, as of the time of writing (November 20, 2025), no sparse retrieval model is listed on the MTEB (Multilingual, v2) leaderboard.

[2] Code and models will be released after notification date.

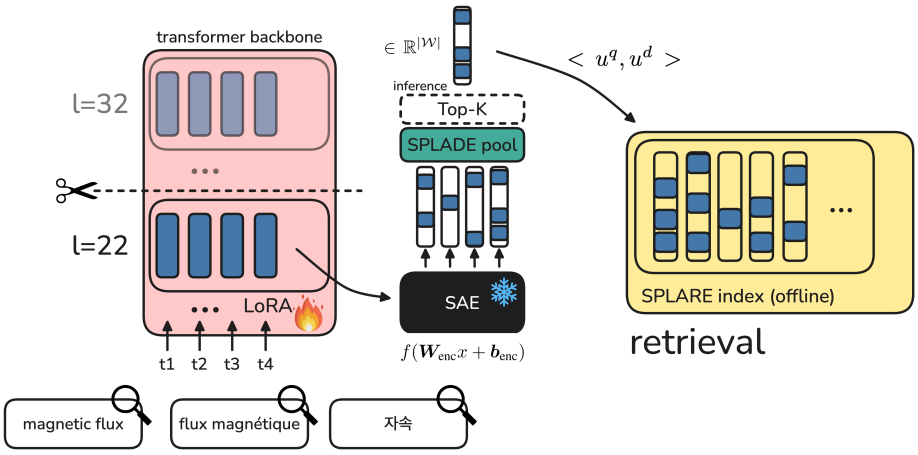

Figure 1: Overview of SPLARE. A pre-trained SAE can be inserted at any layer l of the LLM to get sparse latent representations of input tokens. These token-level representations are then aggregated into a single sparse vector using a pooling mechanism analogous to SPLADE. During training, we only fine-tune the LLM parameters (via LoRA adapters) while keeping the SAE frozen.

He et al., 2024; Cunningham & Conerly, 2024; Deng et al., 2025). Large sparse autoencoders are also notoriously hard and costly to train. Recently, high-quality large scale open-source SAEs have became available to the research community. In particular, we rely in this work on the Llama Scope series of models He et al. (2024) which offers SAEs trained on Llama-3.1-8B and the Gemma Scope suite Lieberum et al. (2024) which offers SAEs trained on Gemma-2-2B, 9B and 27B models.

2.2 SPLADE

Learned Sparse Retrieval (LSR) models aim to encode input sequences into high-dimensional sparse representations. Among these approaches, the SPLADE family of approaches (Formal et al., 2021; 2022a; Lassance et al., 2024) has emerged as the state-of-the-art method, achieving performance comparable to or exceeding that of dense embedding models in many settings. Given an input sequence tokenized as $t = (t_1, t_2, \ldots, t_n)$ and fed through all the layers of the transformer, SPLADE generates a sequence of logits $(v_1, v_2, \ldots, v_n)$ by projecting each final hidden state $(h_1, h_2, \ldots, h_n)$ onto the vocabulary space $\mathcal{V}$ using the language modeling head, i.e., via a linear transformation based on the token embedding matrix. The weights $(v_{ij})_{j \in \mathcal{V}}$ correspond to an unnormalized log-probability distribution over $\mathcal{V}$ for token t_i, where each output dimension j is actually associated with the token it represents. To obtain a single sequence-level representation, SPLADE first applies a term saturation function, before max-pooling over the sequence:

$$u_j = \max_{i=1\ldots n} \log\left(1 + \text{ReLU}(v_{ij})\right), j \in \mathcal{V} \tag{2}$$

Given these sparse representations $u \in \mathbb{R}^{|\mathcal{V}|}$ for queries and documents, relevance scores are computed as a sparse dot product $s(q,d) =< u^q, u^d >$. This operation can be efficiently supported using inverted index structures together with specialized query processing techniques (Tonellotto et al., 2018; Bruch et al., 2024c; Zobel & Moffat, 2006).

3 METHOD

3.1 SPLARE

Conceptually, SPLARE closely parallels SPLADE but operates in the latent representation space. Rather than projecting the final hidden states of the language model onto the vocabulary space via the LM head, SPLARE employs sparse autoencoders to transform representations from a selected layer into a sparse latent space, which can be interpreted as a latent vocabulary.

Let $(\boldsymbol{W}_{\text{dec}}, \boldsymbol{b}_{\text{dec}})$ in Eq. 1 denote the SAE's encoder parameters at a given layer l of the transformer[3]. Similary to SPLADE, we can obtain sequences of sparse latent logits $(w_1, w_2, \ldots, w_n)$ by mapping the hidden states at layer l with the SAE encoder as illustrated in Figure 1. The weights $(w_{ij})_{j \in \mathcal{W}} \in \mathbb{R}^{|\mathcal{W}|}$ contain the sparse list of latent features associated with token i in the input sequence. It can be used in place of the vocabulary decomposition to compute sequence-level representations for input queries or documents into a sparse set of latent features using the same type of pooling mechanism as in Eq. 2—which we refer to as SPLADE-pool in Figure 1.

3.2 Training

Training LSR Models The training procedure for LSR models mirrors that of dense embedding models. While contrastive learning Oord et al. (2018); Chen et al. (2020) is the de-facto approach to train state-of-the-art dense models Lee et al. (2025b); Zhang et al. (2025), we instead adopt a distillation-based approach using a cross-encoder teacher model Nogueira & Cho (2020) to train our sparse embeddings. Distillation is a common toolbox to train retrieval models (Hofstätter et al., 2020; Lin et al., 2020), but has been overlooked in the context of LLM-based embeddings. Specifically, we optimize the Kullback–Leibler divergence between the teacher and student relevance distributions (Lin et al., 2020). Given a query q, $(d_1, d_2, \ldots d_m)$ which contains a positive document and a pool of hard negatives, $(s_1, s_2, \ldots s_m)$ the corresponding teacher scores for documents d_i with respect to q, and τ a temperature parameter, the training loss is given by:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{m} p_i \left(\log p_i - \log \hat{p}_i \right), \quad \hat{p}_i = \frac{e^{s(q,d_i)/\tau}}{\sum_j e^{s(q,d_j)/\tau}}, \quad p_i = \frac{e^{s_i}}{\sum_j e^{s_j}} \tag{3}$$

Sparsity To encourage sparsity in query and document representations, LSR models are typically trained with a sparsity-inducing regularization term, analogous to that used in SAEs. Following Porco et al. (2025), we adopt a slight modification of the original FLOPS loss Paria et al. (2020) employed in SPLADE. The final loss is $\mathcal{L} = \mathcal{L}_{\text{KL}} + \lambda_q \ell^q_{\text{DF-FLOPS}} + \lambda_d \ell^d_{\text{DF-FLOPS}}$.

The sparsity of LSR approaches plays a crucial role in determining both effectiveness and computational efficiency on retrieval benchmarks. However, the sparsity induced by $\mathcal{L}$ can vary significantly depending on the model configuration, backbone architecture, SAE suite, and dataset characteristics. Achieving a desired target sparsity would require continuous adjustment of $\lambda_{d,q}$. To mitigate this challenge and establish a more robust training setup, we additionally apply Top-K pooling *at inference time*, as illustrated in Figure 1. This strategy allows us to train a single model with moderate sparsity—using fixed, conservative values of $\lambda_{d,q}$—while systematically studying the effect of pooling without the need for re-training. Although some prior works have entirely replaced explicit sparsity regularization with Top-K pooling (Lassance et al., 2023; Doshi et al., 2024), our initial experiments with this approach yielded inferior results. Finally, we note that while SPLARE is initialized with an SAE—which produces inherently sparse token-level representations—sequence-level sparsity at initialization remains relatively high (e.g., a few thousands non-zero values). As a result, additional sparsity regularization is required to ensure the model achieves the desired efficiency. It is also worth noting that LSR models are usually hard to train and require a careful initialization of the projection head. While the LM head or a SAE can provide a suitable initialization, training an LSR model entirely from scratch is highly difficult and consistently results in lower performance.

4 Experimental setup

Training Data We conduct two large sets of experiments: § 5 contains various ablations and analyses for models trained on English data on the MS MARCO dataset Bajaj et al. (2018). In § 6, we further extend training to a larger set of publicly available data, including multi-lingual datasets. We do not prepend any special instructions or prefix to our input sequences—which could only likely yield further improvements. To ease reproducibility, we also refrain from any form of pre-finetuning or synthetic data generation Lee et al. (2025b); Günther et al. (2025); Zhang et al. (2025), both of which have recently become common practice for achieving top results on the MTEB benchmark. We detail in Appendix A our two training settings.

[3]Note that we only rely on the encoder parameters, as we only aim to extract sparse features from representations. Also note that we consider SAEs trained on the residual streams of the transformer.

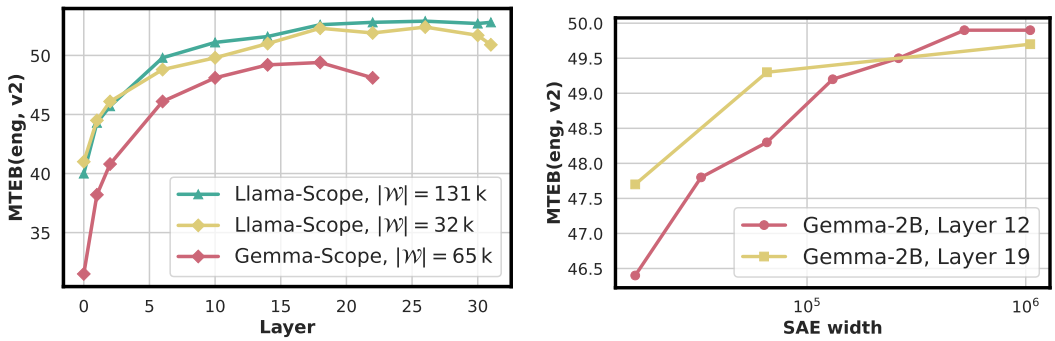

Figure 2: *(Left)* Performance across layers on Llama Scope (Llama-3.1-8B) and Gemma Scope (Gemma-2-2B). *(Right)* Performance with increasing SAE width on Gemma-2. Evaluation done with Top-K $= (40, 400)$.

Evaluation MTEB Muennighoff et al. (2023) and MMTEB Enevoldsen et al. (2025) are the most widely adopted benchmarks for evaluating embedding models. Our evaluation focuses only on the *retrieval* subsets of these benchmarks, excluding other task categories. In addition to the English and Multilingual splits, we also report results on domain-specific subsets of MTEB, including Code, Medical, Law, and Chemical domains. Given SPLARE's strong performance in multilingual settings, we further place particular emphasis on this aspect by including language-specific splits of MMTEB for five languages, as well as evaluations on the MIRACL Zhang et al. (2023) and XTREME-UP Ruder et al. (2023) datasets. The latter introduces a challenging cross-lingual retrieval task, requiring retrieval from an English corpus using queries from low-resource languages. We also report results on MS MARCO Bajaj et al. (2018) and BEIR Thakur et al. (2021) (Appendix C).

While our approach is broadly applicable to any pre-trained SAE, we conduct the majority of our experiments using the Llama Scope model suite He et al. (2024), built on Llama-3.1-8B (et al., 2024). During training, we fine-tune the backbone with LoRA adapters Hu et al. (2022) while keeping SAE parameters frozen. Preliminary experiments indicated that this strategy not only improves performance but also simplifies training. Moreover, it preserves the interpretability of the latent feature space Lin (2023). As in prior work (Zeng et al., 2025; BehnamGhader et al., 2024; Lei et al., 2025), we enable bidirectional attention across all backbones and pretrain them with Masked Next Token Prediction. Following the exact procedure of Zeng et al. (2025), we mask 20% of tokens in the MS MARCO corpus and train for $10k$ steps which takes about five hours. Bidirectional attention is particularly important for LSR models since pooling occurs at every position of the input sequence, unlike dense models that rely on the <EOS> token. Full details of our experimental hyperparameters are provided in Appendix B. Unless stated otherwise, retrieval evaluation is performed using Top-K pooling, with default values of $k = 40$ for queries and $k = 400$ for documents. For our multilingual models (§ 6), we additionally rely on model averaging (Wortsman et al., 2022) from several training runs, which boosts generalization performance (Lee et al., 2025b; Zhang et al., 2025).

We are mainly interested in ***comparing SPLARE to current state-of-the-art LSR methods, which are all vocabulary-based***. To this end, we perform controlled comparisons with a SPLADE model built on the same Llama-3.1-8B backbone—following the methodology of (Doshi et al., 2024; Zeng et al., 2025)—and trained under identical settings. We refer to this baseline as SPLADE-Llama.

5 Analysis and Design Choices for SPLARE Models

We first conduct a series of ablation studies in a controlled, English-only setting. At this stage, our primary objective is to compare SPLARE's latent representations with traditional vocabulary-based approaches (i.e., our SPLADE-Llama baseline). Specifically, we aim to address the following research questions: *(i)* At which transformer layer depth do we obtain the most effective sparse latent representations for retrieval? *(ii)* How does the width of the SAE affect retrieval performance? *(iii)* What are the efficiency–effectiveness trade-offs introduced by the latent vocabulary? *(iv)* Do the sparse latent features learned by the SAE yield improvements over equivalent SPLADE models?

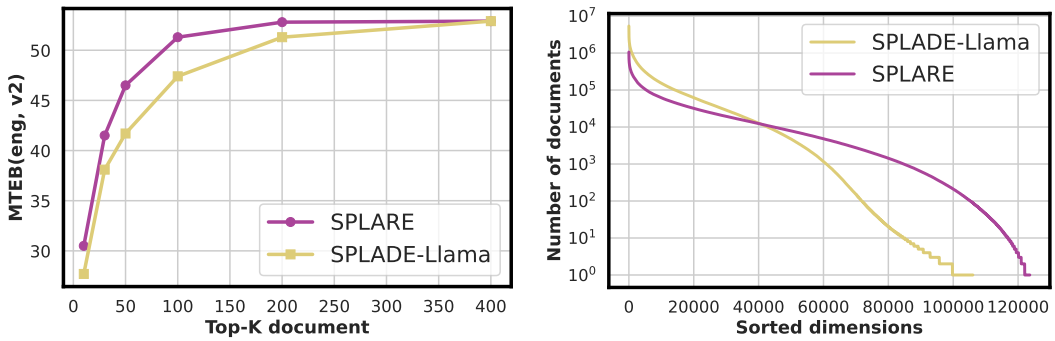

Figure 3: *(Left)* Impact of pruning documents with Top-K (with $k = 40$ for queries). *(Right)* MS MARCO index distribution for SPLARE and SPLADE ($8.8M$ documents).

Performance and Layer Depth We train SPLARE models at varying depths on Llama-3.1-8B, using SAEs from Llama Scope with two widths $|\mathcal{W}| \in \{32k, 131k\}$, and on Gemma-2-2B, using Gemma Scope with width $|\mathcal{W}| = 65k$, and report the average MTEB (English, v2) performance in Figure 2 (*Left*). Interestingly, the highest performance is consistently achieved at about two-thirds of the model depth, i.e., around layer 20 (out of 32) for Llama Scope and 16 (out of 26) for Gemma Scope. These findings are consistent with prior work suggesting that intermediate transformer layers often yield richer representations for retrieval tasks (Skean et al., 2025; Zhuang et al., 2025; Wang et al., 2025). A further advantage of using intermediate layers is the reduction in retriever size and, consequently, inference latency—an improvement over SPLADE models, which require processing through all layers of the LLM (see Appendix F). For the remainder of the paper, our main SPLARE models are trained at layer 26 of Llama-3.1-8B, yielding a 7B-parameter model (including the SAE parameters).

How does the width of the SAE affect retrieval performance? Unlike SPLADE models, the dimensionality of SPLARE's feature space—determined by the SAE width $|\mathcal{W}|$—is not constrained by the LLM's vocabulary size. To study the effect of SAE width on retrieval effectiveness, we train multiple SPLARE models using Gemma Scope, which offers a broader range of SAE configurations. Especially, we consider SAEs at layers 12 and 19 of Gemma-2-2B with widths $|\mathcal{W}| \in \{2^{14} \approx 16k, 2^{15}, \ldots, 2^{20} \approx 1M\}$. We report the resulting average MTEB (English, v2) performance in Figure 2 (*Right*). Our results show a roughly log-linear relationship between SAE width and retrieval effectiveness, providing a scaling mechanism for improved performance—something not possible with SPLADE's fixed vocabulary size. Prior work has shown that SAEs can scale to widths as large as $14M$ on very large LLMs (Templeton et al., 2024), though such models remain proprietary. Llama Scope, while limited to $|\mathcal{W}| \in \{32k, 131k\}$, exhibits the same scaling effect consistently across layers (Figure 2, *(Left)*). These experiments also highlight that the approach is transferable across different backbone architectures. Despite the availability of much wider SAEs in Gemma Scope, we observe that Llama Scope models achieve superior overall performance. Consequently, we report results using this model (with $|\mathcal{W}| = 131k$) for all subsequent experiments.

Effectiveness–efficiency trade-off Sparse retrieval methods achieve efficiency through the use of dedicated inverted index structures and exact (Zobel & Moffat, 2006; Tonellotto et al., 2018) or approximate (Bruch et al., 2024a) query processing algorithms. In all cases, obtaining highly sparse representations is critical for achieving low-latency retrieval. While SPLADE has been successfully adapted to LLM backbones, efficiency considerations have generally been overlooked. As discussed in § 3.2, LSR models can easily become "dense" in practical scenarios, which undermines their efficiency.

We study the relationship between SPLARE performance and sparsity by capping, at inference time, the number of activated features for documents vectors using Top-K pooling. Results are shown in Figure 3 (*Left*). SPLARE exhibits substantially greater robustness to document pruning: when indexing only Top-K $= 100$ document features, its performance drops by merely 2%, compared to over 6% for SPLADE. This difference can be partially attributed to SPLARE's more compact and structured latent feature space as well as the fact that SPLADE models based on LLMs are inher-

Table 1: Average performance on various MTEB splits. English models are trained on MS MARCO only (§ 5). Multilingual models are trained on a large-scale multilingual training set (§ 6). Evaluation done with Top-K $= (40, 400)$.

	English	Multilingual	Code	Medical	Law	ChemTEB
English Models						
SPLADE-v3 (Lassance et al., 2024)	50.7	38.1	44.5	44.2	40.4	75.6
Lion-SP-8B (Zeng et al., 2025)	48.5	50.0	53.3	54.4	48.5	71.1
SPLADE-Llama	**52.9**	54.3	**57.3**	61.0	49.0	**75.9**
SPLARE	**52.9**	**56.3**	55.1	**62.9**	**51.2**	70.0
Multilingual Models						
SPLADE-Llama	58.4	60.3	**63.6**	67.1	57.5	75.7
SPLARE	**58.6**	**60.9**	60.7	**68.0**	**58.1**	**77.2**

ently harder to sparsify. As we show in Appendix E, this difference translates into lower query latency at a given accuracy level, when evaluated using Seismic (Bruch et al., 2024a;b; 2025). For reference, performing retrieval with SPLARE (Top-K $= (40, 400)$) on MS MARCO ($8.8M$ documents) requires only about $5ms$ per query only—without accounting model inference. Figure 3 *(Right)* further illustrates the distributions of activated features after training. Notably, SPLARE utilizes a much larger portion of the available feature space, activating nearly all dimensions, in contrast to SPLADE, which relies on fewer than $100k$ dimensions (out of $128k$). Moreover, SPLARE exhibits a more balanced activation distribution across features. By comparison, SPLADE tends to over-activate a small subset of dimensions (Mackenzie et al., 2023; Lei et al., 2025).

Comparison of lexical and latent features Finally, we compare the performance of SPLARE with existing top LSR methods trained on the English MS MARCO dataset. In particular, Lion-SP-8B Zeng et al. (2025) represents the most effective contemporary SPLADE adaptation for LLM-based retrieval. We show the results on Table 1 for various splits of MTEB (English Models). First, notice that SPLADE-Llama (our baseline) already outperforms Lion-SP-8B. We further observe that SPLARE consistently outperforms competing methods on both multilingual and several out-of-domain evaluation sets. In particular, it achieves an improvement of roughly two points on the multilingual split and shows superior performance on the Law and Medical retrieval benchmarks—though its advantage diminishes on the Code and Chemical splits. The observed multilingual generalization from English-only training is unsurprising, given the language-agnostic nature of SAE features. With respect to out-of-domain performance, we hypothesize that the decomposition mechanism of SAEs transfers more effectively across domains, whereas SPLADE-like models rely on explicit in-domain training to adequately expand their vocabulary representations. Meanwhile, the performance drop on the Code tasks is likely due to the highly domain-specific nature of code retrieval, which does not align well with the features learned by the SAE (a trend that is further supported by our observations in § 6). To illustrate this behavior, we provide in Appendix G (Figures 15-17) several examples where SPLARE underperforms compared to SPLADE on MTEB Code. In these cases, the top activated features appear overly generic rather than specialized to code semantics. This suggests that for highly domain-specific scenarios such as code retrieval, dedicated SAEs trained on code-focused corpora may be more appropriate. We leave this direction for future work.

6 Multi-Lingual Models

In § 5, we showed how the latent feature space of the SAE offers some advantages for LSR models—when compared to the vocabulary space—*in a controlled English-based setting*. In § 6.1, we further extend those findings for multilingual models, by training models on a large-scale multilingual dataset and broadening the evaluation to cover a more diverse set of benchmarks, as detailed in § 3.2. In § 6.2, we compare SPLARE to concurent models on (M)MTEB and XTREME-UP.

Table 2: Multilingual comparison of SPLARE and SPLADE (Top-K $= (40, 400)$).

	indic	sca	deu	fra	kor	XTREME-UP	MIRACL
SPLADE-Llama	90.1	70.4	55.4	65.6	73.7	56.3	67.9
SPLARE	**91.2**	70.4	**56.2**	65.6	**74.9**	**59.8**	69.6

6.1 Comparing Latent Models to Lexicon-based approaches

Table 1 (Multilingual models) compares the average performance of multi-lingual SPLARE and SPLADE across the various MTEB splits, with full results provided in Appendix D. Overall, SPLARE consistently outperforms its vocabulary-based counterpart, with the exception of the Code split. A closer inspection of individual datasets within the Multilingual split reveals that SPLARE systematically outperforms SPLADE, a trend further confirmed by Table 9, which highlights the superior performance of latent-based LSR models in multilingual settings. On XTREME-UP, SPLARE also maintains its performance advantage. Comprehensive results for both MIRACL and XTREME-UP, along with comparisons to concurrent approaches, are provided in Appendix D. Notably, SPLARE exhibits particularly strong results on the hidden test sets of MIRACL (Table 10) and the low-ressource languages of XTREME-UP (see also Table 3).

6.2 Comparing to Top Models

Finally, we compare SPLARE to top models from the MTEB leaderboard in Table 3. SPLARE reaches an average score of 60.9 (for the pooled version), ***making it among the top 10 models on MTEB(Multilingual, v2) retrieval and the top-1 LSR model***. Notably, these results are achieved without relying on private or synthetic data and without any pre-finetuning. This is also particularly interesting, as open models like gte-Qwen2-7B instruct or NV-Embed-v2 rely on 3584-d (resp. 4096-d) dense vectors to encode queries and documents, while SPLARE* only needs 40 features (resp. 400) to encode queries (resp. documents) in its high-dimensional feature space. We also observe an average gain of $+1$ point for the non-pooled version, albeit at the cost of higher retrieval complexity. On the other hand, extremely sparse models (Top-K $= (10, 100)$) still offer competitive performance. Note that in practical retrieval scenarios, dense embeddings often require dimensionality-reduction techniques Kusupati et al. (2022) and/or approximate nearest-neighbor search algorithms Johnson et al. (2019) algorithms—whose performance degradation is rarely reported on standard benchmarks. In contrast, sparse retrieval methods natively support efficient exact search without incurring such compromises. Finally, we also report results for a SPLARE model trained at layer 6 (SPLARE-2B). Although its performance is somewhat lower than that of the full SPLARE model (7B parameters), it remains strong—particularly on the XTREME-UP dataset. Importantly, this model is substantially more efficient and therefore offers a different, and often attractive, point on the effectiveness–efficiency trade-off curve.

6.3 Interpretability: insight into SPLARE mechanics

Finally, we provide interpretability insights for SPLARE. We leverage Neuronpedia (Lin, 2023) to obtain explanations for individual SAE features—which, as a reminder, remain frozen during fine-tuning—and list the top features contributing to a document's relevance with respect to a given query. For SPLADE, by contrast, we report the tokens with the highest relevance contributions. Figure 4 illustrates a cross-lingual example from XTREME-UP from Tamil to English. The features activated by SPLARE align well with meaningful concepts present in both the query and document. They correspond to coherent, language-agnostic concepts which combine into a comprehensive description of the data point. In contrast, SPLADE exhibits a higher degree of redundancy (e.g., separate activations for "Indian" and "indian") and predominantly relies on Latin-script tokens—effectively defaulting to English subword representations—which provide less informative signals in this setting. Further examples are given in Appendix G.

Table 3: Average MTEB retrieval performance of SPLARE (Multilingual) against top models. Multilingual (resp. Eng) refers to (Multilingual, v2) (resp. MTEB(eng, v2)). As of November 20, 2025, SPLARE* ranks in the top-10 models on MTEB(Multilingual, v2) retrieval. For XTREME-UP (MRR@10), we report results from (Lee et al., 2025b). Unless specified, evaluation for SPLARE is done with Top-K $= (40, 400)$.

	Eng	Multilingual	XTREME-UP
Top Open Source models			
e5-mistral-7b-instruct (Wang et al., 2024a)	57.6	55.8	-
NV-Embed-v2 (Lee et al., 2025a)	62.8	56.7	-
multilingual-e5-large-instruct (Wang et al., 2024b)	53.5	57.1	18.7
GritLM-7B (Muennighoff et al., 2024)	55.0	58.3	-
SFR-Embedding-Mistral (Meng et al., 2024)	59.3	59.4	-
Linq-Embed-Mistral (Kim et al., 2024)	60.1	58.7	24.6
gte-Qwen2-7B-instruct (Li et al., 2023b)	58.1	60.1	17.4
voyage-3-large (AI, 2025)	53.5	66.1	39.2
jina-embeddings-v4 (Günther et al., 2025)	56.2	66.4	-
inf-retriever-v1 (Yang et al., 2025)	64.1	66.5	-
Qwen-3-Embedding-8B (Zhang et al., 2025)	69.4	70.9	-
Commercial models			
Cohere-embed-multilingual-v3.0 (Cohere, 2023)	55.7	59.2	-
text-embedding-3-large (OpenAI, 2024)	58.0	59.3	18.8
gemini-embedding-001 (Lee et al., 2025b)	64.4	67.7	64.3
SPLARE*	58.6	60.9	59.8
SPLARE, no-pooling	59.8	61.9	61.7
SPLARE, Top-K $= (20, 200)$	55.9	59.3	55.2
SPLARE, Top-K $= (10, 100)$	50.7	56.2	48.6
SPLARE-2B	55.5	57.6	42.7

Figure 4: Retrieval example from XTREME-UP: `Tamil` → `English`

`Query`: அங்கிலேயர்கள் ஆட்சியில் சராசரியாக எத்தனை இந்தியர்கள் இறந்தனர்

`Translation`: On average, how many Indians died under British rule?

`Positive document`: Indian Army during World War II: The British Indian Army fought in Ethiopia against the Italian Army, in Egypt, Libya, Tunisia and Algeria against both the Italian and German Army, and, after the Italian surrender, against the German Army in Italy. [...]

SPLARE | top features (doc rank = 4) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 23) Token	%
elements related to historical or cultural contexts	10.5	␣Indian	12.6
mentions of India and its relation to various contexts	8.7	␣Indians	11.2
descriptions that contrast traditional experiences with unique local accommodation	7.5	␣casualties	9.0
mentions of colonial powers, specifically Britain and France	6.6	␣India	8.5
references to military casualties and losses	6.5	␣indian	7.6
quantitative statistics and casualties related to wars and conflicts	5.9	␣British	7.5
information related to economic data and connectivity issues in India	5.2	␣deaths	6.8
references to protests and civil rights movements	5.0	␣fatalities	5.3
references to historical events and political movements	4.9	␣india	4.7
references to corporate structure and business dynamics	4.4	␣Raj	4.5

7 Related Works

LLMs and Retrieval Dense embedding models derived from LLMs have demonstrated substantial gains over traditional BERT-style encoders (Lee et al., 2025b; Zhang et al., 2025). Recent approaches such as LLM2Vec BehnamGhader et al. (2024) or GritLM Muennighoff et al. (2024) highlight how LLMs can be effectively adapted into powerful text encoders by incorporating bidirectional attention. Beyond providing stronger backbone architectures, LLMs have also significantly advanced retrieval model training, enabling the generation of high-quality synthetic data and improved filtering of training samples (Wang et al., 2024a; Lee et al., 2025a;b; Zhang et al., 2025; Dai et al., 2023). Nonetheless, despite the impressive progress of dense embeddings, con-

trolled evaluations have shown that they can still be outperformed by alternative architectures such as multi-vector models or sparse retrievers (Zeng et al., 2025; Faysse et al., 2025; Chen et al., 2024a).

Sparse Autoencoders and Retrieval Sparse autoencoders have primarily been employed in Information Retrieval (IR) to approximate dense representations for efficient nearest-neighbor search. Given a dense embedding model, these approaches learn to map query and document vectors into sparse latent representations that preserve the structure of the original embedding space (Lassance et al., 2021; Borges et al., 2023; Park et al., 2025; Kang et al., 2025; Wen et al., 2025). SAEs have also been used to interpret dense emebddings in both IR O'Neill et al. (2024) and Recommender Systems (Kasalický et al., 2025; Klenitskiy et al., 2025). Most closely related to our work is (Park et al., 2025), which shows that SAE-derived features can serve as effective indexing units. However, all prior studies train SAEs on top of an *already-trained dense retriever*. In contrast, our approach leverages pre-trained SAEs on the base LLM and fine-tunes an LSR model directly in a SPLADE-like fashion, allowing for tighter integration of relevance and sparsity when training the sparse representations.

8 Conclusion

In this work, we investigated two complementary research directions: Sparse autoencoders and Learned Sparse Retrieval models. We demonstrated that SAEs provide a natural foundation for LSR by yielding semantically rich and multilingual latent features that overcome the vocabulary dependence of traditional LSR approaches. Our experiments show that SAE-based LSR models consistently outperform vocabulary-based counterparts, particularly in multilingual and out-of-domain scenarios. Finally, we introduced SPLARE, a competitive 7B-parameter multilingual model capable of producing generalizable sparse latent embeddings, thereby paving the way for more robust, versatile, and cross-lingual retrieval across diverse domains and modalities.

References

Voyage AI. Voyage-3 large. https://blog.voyageai.com/2025/01/07/voyage-3-large/, January 2025. Accessed: 2025-09-24.

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.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Luís Borges, Bruno Martins, and Jamie Callan. Kale: Using a k-sparse projector for lexical expansion. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, pp. 13–22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700736. doi: 10.1145/3578337.3605131. URL https://doi.org/10.1145/3578337.3605131.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 152–162, 2024a.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. Pairing clustered inverted indexes with κ-nn graphs for fast approximate retrieval over learned sparse representations. In *Proceedings of the 33rd International ACM Conference on Information and Knowledge Management (CIKM)*, pp. 3642–3646. ACM, 2024b. doi: 10.1145/3627673.3679977. URL https://doi.org/10.1145/3627673.3679977.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 152–162. ACM, 2024c. doi: 10.1145/3626772.3657769. URL https://doi.org/10.1145/3626772.3657769.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, Rossano Venturini, and Leonardo Venuta. Investigating the scalability of approximate sparse retrieval algorithms to massive datasets. In *Advances in Information Retrieval*, pp. 437–445. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-88714-7_43. URL https://doi.org/10.1007/978-3-031-88714-7_43.

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

Jianlyu 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. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Cohere. Introducing embed v3. https://cohere.com/blog/introducing-embed-v3, November 2023. Accessed: 2025-09-24.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track, 2021. URL https://arxiv.org/abs/2102.07662.

Hoagy Cunningham and Tom Conerly. Comparing topk and gated saes to standard saes. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/june-update/index.html#topk-gated-comparison.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gmL46YMpu2J.

Hervé Déjean, Stephane Clinchant, Carlos Lassance, Simon Lupart, and Thibault Formal. Benchmarking middle-trained language models for neural search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 1848–1852, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591956. URL https://doi.org/10.1145/3539618.3591956.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Meet Doshi, Vishwajeet Kumar, Rudra Murthy, Vignesh P, and Jaydeep Sen. Mistral-splade: Llms for better learned sparse retrieval, 2024. URL https://arxiv.org/abs/2408.11119.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. A thorough comparison of cross-encoders and llms for reranking splade, 2024. URL https://arxiv.org/abs/2403.10407.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zl3pfz4VCV.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2288–2292, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463098. URL https://doi.org/10.1145/3404835.3463098.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2353–2359, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531857. URL https://doi.org/10.1145/3477495.3531857.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Match your words! a study of lexical matching in neural information retrieval. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pp. 120–127, Berlin, Heidelberg, 2022b. Springer-Verlag. ISBN 978-3-030-99738-0. doi: 10.1007/978-3-030-99739-7_14. URL https://doi.org/10.1007/978-3-030-99739-7_14.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval, 2025. URL https://arxiv.org/abs/2506.18902.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders, 2024. URL https://arxiv.org/abs/2410.20526.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Hao Kang, Tevin Wang, and Chenyan Xiong. Interpret and control dense retrieval with sparse latent features. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 700–709, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.58. URL https://aclanthology.org/2025.naacl-short.58/.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rNfzT8YkgO.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Petr Kasalický, Martin Spišák, Vojtěch Vančura, Daniel Bohuněk, Rodrigo Alves, and Pavel Kordík. The future is sparse: Embedding compression for scalable retrieval in recommender systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pp. 1099–1103, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713644. doi: 10.1145/3705328.3748147. URL https://doi.org/10.1145/3705328.3748147.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog, 2024. URL https://getlinq.com/blog/linq-embed-mistral/.

Anton Klenitskiy, Konstantin Polev, Daria Denisova, Alexey Vasilev, Dmitry Simakov, and Gleb Gusev. Sparse autoencoders for sequential recommendation models: Interpretation and flexible control, 2025. URL https://arxiv.org/abs/2507.12202.

Weize Kong, Jeffrey M. Dudek, Cheng Li, Mingyang Zhang, and Michael Bendersky. Sparseembed: Learning sparse lexical representations with contextual embeddings for retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2399–2403, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592065. URL https://doi.org/10.1145/3539618.3592065.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, December 2022.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Carlos Lassance. Extending english ir methods to multi-lingual ir, 2023. URL https://arxiv.org/abs/2302.14723.

Carlos Lassance, Thibault Formal, and Stéphane Clinchant. Composite code sparse autoencoders for first stage retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2136–2140, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463066. URL https://doi.org/10.1145/3404835.3463066.

Carlos Lassance, Simon Lupart, Hervé Déjean, Stéphane Clinchant, and Nicola Tonellotto. A static pruning study on sparse neural retrievers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 1771–1775, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591941. URL https://doi.org/10.1145/3539618.3591941.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. Splade-v3: New baselines for splade. *arXiv preprint arXiv:2403.06789*, 2024.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=lgsyLSsDRe.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025b. URL https://arxiv.org/abs/2503.07891.

Yibin Lei, Tao Shen, Yu Cao, and Andrew Yates. Enhancing lexicon-based text embeddings with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18986–19001, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.930. URL https://aclanthology.org/2025.acl-long.930/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. Making large language models a better foundation for dense retrieval, 2023a.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. Making text embedders few-shot learners. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wfLuiDjQ0u.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023b. URL https://arxiv.org/abs/2308.03281.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*, 2020.

Simon Lupart, Thibault Formal, and Stéphane Clinchant. Ms-shift: An analysis of ms marco distribution shifts on neural retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, pp. 636–652, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-28243-0. doi: 10.1007/978-3-031-28244-7_40. URL https://doi.org/10.1007/978-3-031-28244-7_40.

Guangyuan Ma, Yongliang Ma, Xuanrui Gou, Zhenpeng Su, Ming Zhou, and Songlin Hu. Lightretriever: A llm-based hybrid retrieval architecture with 1000x faster query inference, 2025. URL https://arxiv.org/abs/2505.12260.

Joel Mackenzie, Andrew Trotman, and Jimmy Lin. Efficient document-at-a-time and score-at-a-time query evaluation for learned sparse representations. *ACM Trans. Inf. Syst.*, 41(4), March 2023. ISSN 1046-8188. doi: 10.1145/3576922. URL https://doi.org/10.1145/3576922.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1723–1727, 2021.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative Representational Instruction Tuning, February 2024. URL http://arxiv.org/abs/2402.09906. arXiv:2402.09906 [cs].

Suraj Nair, Eugene Yang, Dawn J. Lawrie, James Mayfield, and Douglas W. Oard. Learning a sparse representation model for neural clir. In *DESIRES*, pp. 53–64, 2022. URL https://ceur-ws.org/Vol-3480/paper-06.pdf.

Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W. Oard. Blade: Combining vocabulary pruning and intermediate pretraining for scaleable neural clir. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 1219–1229, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591644. URL https://doi.org/10.1145/3539618.3591644.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. A unified framework forlearned sparse retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pp. 101–116, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-28240-9. doi: 10.1007/978-3-031-28241-6_7. URL https://doi.org/10.1007/978-3-031-28241-6_7.

Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. Multimodal learned sparse retrieval with probabilistic expansion control. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, pp. 448–464, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-56059-0. doi: 10.1007/978-3-031-56060-6_29. URL https://doi.org/10.1007/978-3-031-56060-6_29.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

Charles O'Neill, Christine Ye, Kartheik G. Iyer, and John F Wu. Towards interpretable scientific foundation models: Sparse autoencoders for disentangling dense embeddings of scientific concepts. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. URL https://openreview.net/forum?id=mPq3R6jdtD.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. text-embedding-3-large and new embedding models. https://openai.com/index/new-embedding-models-and-api-updates/, January 2024. Accessed: 2025-09-25.

Biswajit Paria, Chih-Kuan Yeh, Ian EH Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. Minimizing flops to learn efficient sparse representations. *arXiv preprint arXiv:2004.05665*, 2020.

Seongwan Park, Taeklim Kim, and Youngjoong Ko. Decoding dense embeddings: Sparse autoencoders for interpreting and discretizing dense retrieval, 2025. URL https://arxiv.org/abs/2506.00041.

Aldo Porco, Dhruv Mehra, Igor Malioutov, Karthik Radhakrishnan, Moniba Keymanesh, Daniel Preoţiuc-Pietro, Sean MacAvaney, and Pengxiang Cheng. An alternative to flops regularization to effectively productionize splade-doc. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 2789–2793, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730163. URL https://doi.org/10.1145/3726302.3730163.

Jingfen Qiao, Thong Nguyen, Evangelos Kanoulas, and Andrew Yates. Leveraging decoder architectures for learned sparse retrieval, 2025. URL https://arxiv.org/abs/2504.18151.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. URL https://arxiv.org/abs/2407.14435.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1856–1884, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.125. URL https://aclanthology.org/2023.findings-emnlp.125/.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Lewis Smith, Senthooran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. Negative results for saes on downstream tasks and deprioritising sae research (gdm mech interp team progress update 2). https://www.alignmentforum.org/posts/4uXCAJNuPKtKBsi28/sae-progress-update-2-draft, 2025.

Livio Soares, Daniel Gillick, Jeremy Cole, and Tom Kwiatkowski. NAIL: Lexical retrieval indices with efficient non-autoregressive decoders. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2574–2589, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.156. URL https://aclanthology.org/2023.emnlp-main.156/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Nicola Tonellotto, Craig Macdonald, Iadh Ounis, et al. Efficient query processing for scalable web search. *Foundations and Trends® in Information Retrieval*, 12(4-5):319–500, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

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

Shuai Wang, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2d matryoshka training for information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 3125–3134, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730330. URL https://doi.org/10.1145/3726302.3730330.

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. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=z19u9B2fCZ.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. Llama nemoretriever colembed: Top-performing text-image retrieval model, 2025a. URL https://arxiv.org/abs/2507.05513.

Zhichao Xu, Aosong Feng, Yijun Tian, Haibo Ding, and Lin Lee Cheong. Csplade: Learned sparse retrieval with causal language models, 2025b. URL https://arxiv.org/abs/2504.10816.

Junhan Yang, Jiahe Wan, Yichen Yao, Wei Chu, Yinghui Xu, and Yuan Qi. inf-retriever-v1 (revision 5f469d7), 2025. URL https://huggingface.co/infly/inf-retriever-v1.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Hansi Zeng, Julian Killingback, and Hamed Zamani. Scaling sparse and dense retrieval in decoder-only llms. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 2679–2684, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730225. URL https://doi.org/10.1145/3726302.3730225.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (eds.), *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 127–137, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.12. URL https://aclanthology.org/2021.mrl-1.12/.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://aclanthology.org/2023.tacl-1.63/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Shengyao Zhuang, Shuai Wang, Fabio Zheng, Bevan Koopman, and Guido Zuccon. Starbucks-v2: Improved training for 2d matryoshka embeddings, 2025. URL https://arxiv.org/abs/2410.13230.

Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38 (2):6–es, July 2006. ISSN 0360-0300. doi: 10.1145/1132956.1132959. URL https://doi.org/10.1145/1132956.1132959.

A EXPERIMENTAL SETTING

We detail below the training sets used for the English and Mutlilingual settings.

English Setting For our ablation study, we restrict training to the MS MARCO dataset, given the computational cost associated with training 7B-parameter models. Our experimental setup closely follows that of SPLADE-v3 (Lassance et al., 2024). For each training query, we mine hard negatives using a SPLADE model and derive distillation targets from reranking scores produced by an open-source DeBERTa-v3 reranker (Déjean et al., 2024). This controlled setting is designed to enable a direct and fair comparison between SPLARE and its vocabulary-based counterpart, SPLADE-Llama.

Multilingual Setting In this more compute-intensive setting, we use the same training set employed for the bge-multilingual-gemma2 model (Li et al., 2025)[4]. This corpus includes several English-centric public datasets (e.g., MS MARCO Bajaj et al. (2018), NQ Kwiatkowski et al. (2019), and HotPotQA Yang et al. (2018)), a large collection of Chinese retrieval datasets, and two multilingual benchmarks: MIRACL Zhang et al. (2023) and Mr.TyDi (Zhang et al., 2021). Since we rely on distillation for training, we only keep samples from this dataset which were annotated using the BGE multilingual reranker Chen et al. (2024b); Li et al. (2023a)[5]. After filtering, the final training set comprises approximately $1.3M$ queries with hard negatives. Notably, some of these datasets correspond to training splits of several MTEB benchmark tasks. While this may constrain the strict evaluation of generalization, this practice has become standard in prior work on general-purpose embedding models (Lee et al., 2025a; Wang et al., 2024a; BehnamGhader et al., 2024).

B HYPER-PARAMETERS

Table 4 gives the hyper-parameters used to train and evaluate SPLARE models and other baselines. Note that the temperature parameters τ is critical and needs to be adapted to each SAE suite. For instance, the optimal τ is different between Llama Scope or Gemma Scope. This depends on the scale of the logits and the initial sparsity of the SAE. For ill-suited τ, it can happen that models actually diverge—for instance, collapse of the ℓ_0. To determine the optimal temperature, we ran a grid search over the values $\{1, 10, 20, 40, 50, 80, 100\}$, and used NanoBEIR[6]'nDCG@10 as an evalution criterion for all models.

SAE choice Gemma-scope contains multiple SAEs for the same layer and width, but with different ℓ_0. In practice, we observed that the initial SAE's ℓ_0 had no critical effect on final performance—most likely because we fine-tune the backbone LLM. We use SAEs with ℓ_0 closest to 100 throughout the paper. Additionally, Llama and Gemma Scope contain residual SAEs as well as MLP and attention stream SAEs. We only used residual SAEs in this paper.

C ENGLISH-ONLY SPLARE FULL RESULTS

We evaluate models from Section 5 (trained on English data only) on several benchmarks, and provide results in Table 1. We show in Table 5 additional evaluation results comparing SPLARE to SPLADE-Llama. We report MRR@10 on MS MARCO Bajaj et al. (2018) and nDCG@10 on TRECDL 19 and TRECDL 20 Craswell et al. (2021) and on all BEIR datasets (Thakur et al., 2021).

D FULL RESULTS

Tables 6—9 provide the full results of several MTEB datasets: English, Multilingual, and various domains and languages.

[4] `hanhainebula/bge-multilingual-gemma2-datadata`

[5] `BAAI/bge-reranker-v2-m3reranker`

[6] `https://huggingface.co/collections/zeta-alpha-ai/nanobeir-66e1a0af21dfd93e620cd9f6`

Table 4: Hyperparameters.

Component	**Value**
LoRA rank r	64
Max training sequence length (english models)	128
Max training sequence length (multilingual models)	256
Epochs	1
Batch size	128
Learning rate	5×10^{-5}
Warmup ratio	0.01
Weight decay	0.
Nb negatives per query	8
λ_d	0.0001
λ_q	0.0001
τ SPLARE (LLama Scope)	80
τ SPLARE (Gemma Scope)	50
τ (SPLADE-LLama)	10
Evaluation max context size	512
Adam βs	0.9, 0.999

Dataset	SPLARE	SPLADE-Llama
arguana	16.0	16.2
climate-fever	18.3	18.0
dbpedia	44.3	44.8
fever	76.0	75.8
fiqa	42.4	42.3
hotpotqa	66.8	67.6
nfcorpus	37.3	36.4
nq	61.6	61.2
quora	87.3	87.9
scifact	72.5	72.9
trec-covid	84.7	82.4
webis	27.2	26.9
scidocs	17.5	17.3
Average	50.2	50.0
MS MARCO (MRR@10)	40.8	40.0
TRECDL 19	77.4	76.3
TRECDL 20	77.3	75.9

Table 5: Full results (nDCG@10 unless specified) on BEIR, MS MARCO and TREC DL for English-based SPLARE and SPLADE-Llama models. Evaluation done with Top-K $= (40, 400)$.

Table 10 compares the SPLARE results on the MIRACL dataset with top multilingual dense retrievers—baseline results are taken from Chen et al. (2024b). On this benchmark, SPLARE obtains an average score of 69.6, only 1.9 points below M3-embeddings (hybrid: dense+sparse) Chen et al. (2024a). Notably, SPLARE is state-of-the-art in English, Finnish, Russian, German and Yoruba, once again indicating its ability to generalize to diverse languages. Note in particular that German and Yoruba are the "secret" languages of MIRACL which were released later *without associated training data*.

E LATENCY MEASURES

We provide per-query retrieval latency as measured on MS MARCO (retrieval from a collection of $8.8M$ documents) for SPLARE and SPLADE-Llama in Figure 5. To measure this, we first index

Task Name	SPLARE	SPLADE-Llama
ArguAna	59.1	64.0
CQADupstackGamingRetrieval	61.6	58.5
CQADupstackUnixRetrieval	44.5	44.1
ClimateFEVERHardNegatives	31.5	38.0
FEVERHardNegatives	89.4	90.4
FiQA2018	53.6	56.4
HotpotQAHardNegatives	77.1	74.0
SCIDOCS	20.4	19.7
TRECCOVID	83.4	81.1
Touche2020Retrieval.v3	65.0	57.6
Average	58.6	58.4

Table 6: Full results of SPLARE and SPLADE-Llama on MTEB(Eng, v2). Evaluation done with Top-K $= (40, 400)$.

Task Name	SPLARE	SPLADE-Llama
AILAStatutes	33.8	34.1
ArguAna	59.1	64.0
BelebeleRetrieval	83.5	82.4
CovidRetrieval	80.6	78.0
HagridRetrieval	98.9	98.6
LEMBPasskeyRetrieval	38.8	38.8
LegalBenchCorporateLobbying	95.3	95.1
MIRACLRetrievalHardNegatives	70.7	68.8
MLQARetrieval	83.2	80.3
SCIDOCS	20.4	19.7
SpartQA	3.6	4.2
StackOverflowQA	86.0	90.2
StatcanDialogueDatasetRetrieval	36.7	32.2
TRECCOVID	83.4	81.1
TempReasonL1	2.4	4.0
TwitterHjerneRetrieval	74.4	75.3
WikipediaRetrievalMultilingual	90.9	89.9
WinoGrande	53.8	48.5
Average	60.9	60.3

Table 7: Full results of SPLARE and SPLADE-Llama on MTEB(Multilingual, v2). Evaluation done with Top-K $= (40, 400)$.

the collection using Seismic Bruch et al. (2024c), and then perform single-threaded retrieval on the saved index. Seeking a very optimal sparse retrieval setup is difficult in general; here we use the very optimized Seismic library without any further tuning. Parameters used to index and retrieve and obtain these latency measurements are given in Table 12.

With the obtained SPLARE models and this simple setup, retrieval takes around 5*ms* per query with maximal accuracy. In low-latency regime (<4*ms*), SPLARE can be used with higher accuracy.

F SPLADE LAYER ABLATION

We showed in § 5 that SPLARE models are usually more effectivte at intermediate layer representations, providing a latency advantage compared to SPLADE. Yet, it is in principle possible to train SPLADE models using intermediate representations as well, by simplify apply the LM head on the intermediate representations. We show results of such a training procedure in Table 13.

Task Name	SPLARE	SPLADE-Llama
ChemTEB		
ChemHotpotQARetrieval	89.3	82.1
ChemNQRetrieval	65.1	69.2
Average	77.2	75.7
Code		
AppsRetrieval	22.6	29.5
COIRCodeSearchNetRetrieval	61.0	72.7
CodeEditSearchRetrieval	72.9	74.4
CodeFeedbackMT	50.3	49.4
CodeFeedbackST	76.3	77.9
CodeSearchNetCCRetrieval	60.4	65.1
CodeSearchNetRetrieval	83.4	86.5
CodeTransOceanContest	81.6	86.6
CodeTransOceanDL	36.3	32.0
CosQA	30.2	30.9
StackOverflowQA	86.0	90.2
SyntheticText2SQL	67.5	67.9
Average	60.7	63.6
Medical		
CUREv1	63.7	56.3
CmedqaRetrieval	28.0	32.2
MedicalQARetrieval	74.8	75.2
NFCorpus	40.0	38.7
PublicHealthQA	85.7	86.0
SciFact	77.8	77.1
SciFact-PL	75.6	73.7
TRECCOVID	83.4	81.1
TRECCOVID-PL	82.8	83.5
Average	68.0	67.1
Law		
AILACasedocs	36.2	36.5
AILAStatutes	33.8	34.1
GerDaLIRSmall	27.5	27.6
LeCaRDv2	62.1	58.6
LegalBenchConsumerContractsQA	86.9	84.6
LegalBenchCorporateLobbying	95.3	95.1
LegalQuAD	55.4	55.2
LegalSummarization	67.8	68.4
Average	58.1	57.5

Table 8: Full results of SPLARE and SPLADE-Llama on MTEB domain specific datasets. Evaluation done with Top-K $= (40, 400)$.

G RETRIEVAL EXAMPLES

We provide in Figures 6—14 multiple examples of scores and explanations obtained for positive documents for some queries on English, Multilingual and multi-domain datasets. We also provide examples on the code domain (Figures 15—17), which highlight some of the limitations on SPLARE on specific domains which might require dedicated SAEs. Notably, in Figure 14 which shows a Tamil example, *the document and query representations coincide for only 6 tokens*, further highlighting SPLADE multilingual limitations. Note that the explanations we used, taken from Neuronpedia, are mostly annotated by LLMs provided with examples of context with features activations. As such, these explanations may remain inaccurate or incomplete.

Task Name	SPLARE	SPLADE-Llama
MTEB(deu, v1)		
GerDaLIR	13.6	13.6
GermanDPR	88.0	85.1
GermanQuAD-Retrieval	95.9	94.9
XMarket	27.2	27.8
Average	56.2	55.4
MTEB(Scandinavian, v1)		
DanFeverRetrieval	41.6	41.5
NorQuadRetrieval	24.7	27.5
SNLRetrieval	98.0	98.3
SweFaqRetrieval	77.9	76.9
SwednRetrieval	82.4	79.9
TV2Nordretrieval	94.2	93.7
TwitterHjerneRetrieval	74.4	75.3
Average	70.5	70.4
MTEB(fra, v1)		
AlloprofRetrieval	56.1	56.9
BSARDRetrieval	66.7	57.7
MintakaRetrieval	47.0	58.6
SyntecRetrieval	90.1	89.1
XPQARetrieval	68.0	65.5
Average	65.6	65.6
MTEB(kor, v1)		
Ko-StrategyQA	83.3	82.4
MIRACLRetrieval	66.6	64.9
Average	74.9	73.7

Table 9: Full results of SPLARE and SPLADE-Llama on MTEB language-specific benchmarks. Evaluation done with Top-K $= (40, 400)$.

Model	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de†	yo†	Avg
Baselines (Prior Work)																			
BM25	39.5	48.2	26.7	7.7	28.7	45.8	11.5	35.0	29.7	31.2	37.1	25.6	35.1	38.3	49.1	17.5	12.0	56.1	31.9
mDPR	49.9	44.3	39.4	47.8	48.0	47.2	43.5	38.3	27.2	43.9	41.9	40.7	29.9	35.6	35.8	51.2	49.0	39.6	41.8
mContriever	52.5	50.1	36.4	41.8	21.5	60.2	31.4	28.6	39.2	42.4	48.3	39.1	56.0	52.8	51.7	41.0	40.8	41.5	43.1
$mE5_{large}$	76.0	75.9	52.9	52.9	59.0	77.8	54.5	62.0	52.9	70.6	66.5	67.4	74.9	84.6	80.2	56.0	56.4	78.3	66.6
$E5_{mistral\text{-}7b}$	73.3	70.3	57.3	52.2	52.1	74.7	55.2	52.1	52.7	66.8	61.8	67.7	68.4	73.9	74.0	54.0	54.1	79.7	63.4
Gemini Embedding	78.3	79.0	58.7	57.0	60.9	78.0	55.6	65.4	54.3	75.1	68.9	73.4	81.0	80.5	80.8	65.7	59.8	88.8	70.1
M3-Emb (Sparse)	67.1	68.9	43.8	38.6	45.1	65.4	35.3	48.2	48.9	56.1	61.5	44.5	57.9	79.1	70.9	36.1	32.5	70.0	53.9
M3-Emb (All)	80.2	81.5	59.6	59.7	63.4	80.4	61.2	63.3	59.0	75.2	72.1	71.7	79.6	88.1	83.7	64.9	59.8	83.5	71.5
SPLADE-Llama	76.9	70.7	57.7	55.6	57.5	78.9	57.1	60.3	57.2	73.0	64.5	71.1	78.7	77.9	78.8	89.8	60.2	56.8	67.9
SPLARE	79.2	72.2	62.0	58.4	59.5	80.5	58.4	62.2	55.5	75.1	66.8	73.8	78.9	75.7	80.8	62.8	61.3	89.1	69.6

Table 10: Multi-lingual retrieval performance on MIRACL dev (nDCG@10). Baseline results are taken from Chen et al. (2024b) and (Lee et al., 2025b). † denotes the two hidden test sets of MIRACL. Evaluation for SPLARE and SPLADE-Llama done with Top-K $= (40, 400)$.

Model	MRR@10
SPLARE	59.8
SPLARE (Eng Only)	41.6
SPLADE-Llama	56.3
SPLADE-Llama (Eng Only)	30.5
Gemini Embedding	64.3
Gemini Embedding (Eng Only)	49.3
Gecko i18n Embedding	35.0
voyage-3-large	39.2
Linq-Embed-Mistral	24.6
multilingual-e5-large-instruct	18.7
gte-Qwen2-7B-instruct	17.4
text-embedding-3-large	18.8

Table 11: XTREME-UP dataset (MRR@10) - Average Scores. Baselines taken from (Lee et al., 2025b). Evaluation for SPLARE done with Top-K $= (40, 400)$.

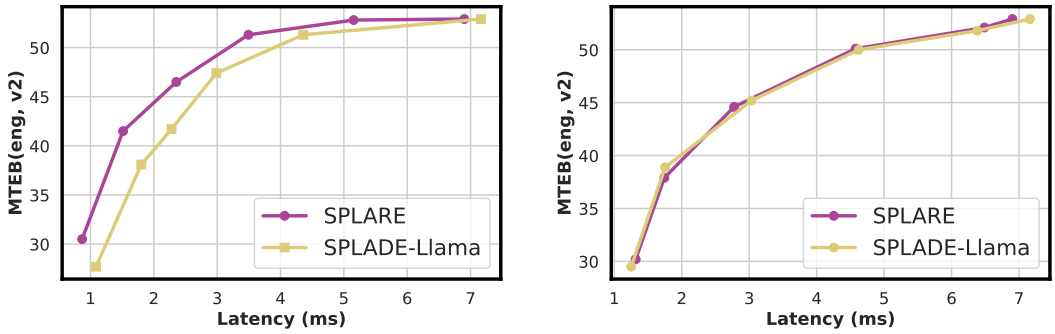

Figure 5: Retrieval Latency (*ms*) when pooling documents *(Left)* or query *(Right)* representations with Top-K. In low-latency settings, SPLARE enables higher accuracy for a given level of latency.

Parameter	Value
k	1000
query_cut	30
heap_factor	0.5
n_knn	0
sorted	False
num_threads	1

Table 12: Seismic retrieval parameters used to measure latency.

SPLADE-Llama at intermediate layers				
Layer No.	18	22	26	31
MTEB(Eng, v2)	0.	43.6	44.5	52.9

Table 13: Training SPLADE-Llama models at intermediate layers leads to strong deterioration. At layer < 22, models collapse during training.

Figure 6: Retrieval example from `BEIR/Scifact`

`Query`: Flexible molecules experience greater steric hindrance in the tumor microenviroment than rigid molecules.
`Positive document`: A solid tumor is an organ composed of cancer and host cells embedded in an extracellular matrix and nourished by blood vessels. A prerequisite to understanding tumor pathophysiology is the ability to distinguish and monitor each component in dynamic studies. Standard fluorophores hamper simultaneous intravital imaging of these components. Here, we used multiphoton microscopy techniques and transge [. . .]

SPLARE | top features (doc rank = 3) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 6) Token	%
references to tumors and their related biological processes	8.6	␣tumor	12.7
terms related to drug delivery and cellular mechanisms	7.5	␣tumors	8.1
terms related to cancer research and metastasis	5.5	␣cancer	7.3
medical conditions and diseases, particularly types of cancer and their characteristics	4.8	␣tum	6.3
concepts related to flexibility in various contexts	4.8	␣nanoparticles	5.3
terms related to cellular processes and immune system functions	4.2	␣Cancer	4.9
references to experimental methods and cell-related terminology	4.0	␣nanop	4.4
terms related to microscopy and micro-level scientific analysis	4.0	␣nan	3.0
variations of the word "tumble" or its related forms	3.6	␣solid	2.6
terms related to cancer and tumors	3.5	␣malignant	2.6

Figure 7: Retrieval example from `BEIR/Scifact`

`Query`: PPAR-RXRs are inhibited by PPAR ligands.
`Positive document`: Heterodimerization is a common paradigm among eukaryotic transcription factors. The 9-cis retinoic acid receptor (RXR) serves as a common heterodimerization partner for several nuclear receptors, including the thyroid hormone receptor (T3R) and retinoic acid receptor (RAR). This raises the question as to whether these complexes possess dual hormonal responsiveness. We devised a strategy to examine . . .

SPLARE | top features (doc rank = 5) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 18) Token	%
terms related to gene transcription regulation	6.2	␣heter	6.2
mathematical variables and expressions	4.3	␣RX	6.0
terms related to dopamine and receptor interactions in the context of medicine and psycho . . .	4.2	rx	5.2
abbreviations or terms related to programming and data structures	4.2	RX	5.0
references to QR codes and VR technologies	4.2	␣receptor	4.9
information about medications used for treating acne	3.8	␣receptors	4.3
references to proteins and their biological functions	3.1	␣nuclear	3.8
terminology related to pharmaceuticals and drug development	2.8	␣Rx	3.7
terms related to cellular functions and regulatory mechanisms	2.8	␣transcription	3.6
terms related to medical and biological concepts, particularly hormones and their effects	2.8	␣rx	3.6

Figure 8: Retrieval example from `BEIR/Climate-Fever`

`Query`: Ocean acidification is the terrifying threat whereby all that man-made CO2 we've been pumping into the atmosphere may react with the sea to form a sort of giant acid bath.
`Positive document`: A greenhouse gas (abbrev . GHG) is a gas in an atmosphere that absorbs and emits radiation within the thermal infrared range . This process is the fundamental cause of the greenhouse effect . The primary greenhouse gases in Earth 's atmosphere are water vapor , carbon dioxide , methane , nitrous oxide , and ozone . Without greenhouse gases , the average temperature of Earth 's surface would be a . . .

SPLARE | top features (doc rank = 6) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 20) Token	%
references to climate change and its associated causes	6.3	␣CO	8.2
statements and discussions regarding climate change-related issues	5.0	␣atmosphere	7.7
terms related to climate change and its impacts	4.4	␣greenhouse	7.3
content related to environmental impacts, particularly concerning carbon dioxide and food . . .	4.1	␣climate	5.7
terms related to carbon emissions and environmental impacts	4.1	␣carbon	5.6
mentions of carbon dioxide and its related metrics or expressions	3.9	␣dioxide	4.7
references to carbon dioxide and its implications in various contexts	3.3	␣anthrop	4.2
references to environmental impact and sustainability	3.2	␣Climate	3.6
references to human activity and its impact on the environment	3.0	␣atmospheric	3.6
mentions of sustainability and environmental impact	3.0	␣Carbon	3.5

Figure 9: Retrieval example from `BEIR/Climate-fever`

`Query`: No state generates as much solar power as California, or has as many people whose jobs depend on it.
`Positive document`: California is the most populous state in the United States and the third most extensive by area . Located on the western (Pacific Ocean) coast of the U.S. , California is bordered by the other U.S. states of Oregon , Nevada , and Arizona and shares an international border with the Mexican state of Baja California . The state capital is Sacramento . Los A . . .

SPLARE | top features (doc rank = 2) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 6) Token	%
references to financial or budgetary topics	7.0	␣California	8.5
references to California	6.8	␣california	6.3
regional references and mentions of cities or places	5.7	␣CA	6.1
references to California and its locations or institutions	4.9	␣Calif	4.9
mentions of political entities and territories	4.3	␣state	4.7
references to political figures and legislation related to California	3.9	␣Californ	4.7
references to geographic locations and regions in California, particularly related to agri . . .	3.8	California	4.6
references to governance, laws, and political contexts	3.3	ifornia	3.6
positive descriptions and references to favorable weather conditions	3.2	␣CAL	3.3
references to California's environmental regulatory bodies and legislation	3.2	␣State	3.2

Figure 10: Retrieval example from `BEIR/Hotpotqa`

`Query`: The Death of Cook depicts the death of James Cook at a bay on what coast?
`Positive document`: Kealakekua Bay is located on the Kona coast of the island of Hawaii about 12 mi south of Kailua-Kona.

SPLARE | top features (doc rank = 3) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 17) Token	%
references to "Bay" or similar geographical features	11.9	␣Hawaii	9.9
references to a specific geographical location or name containing "Bay."	10.7	␣bay	9.7
geographical features and safe navigation routes	9.5	␣Hawai	9.1
references to health and community support systems	9.5	␣Bay	8.2
historical references and significant events	9.3	␣Ke	7.9
references to coastal regions and their characteristics	7.6	Ke	5.4
references to historical sites and landmarks	5.4	Bay	5.1
references to specific geographical locations and their significance in the context of li . . .	5.2	␣Hawaiian	5.0
information related to marine and coastal ecosystems	4.7	bay	4.5
references to sailing, ships, and boating experiences	4.2	␣Haw	4.0

Figure 11: Retrieval example from `MIRACL/Swahili`

`Query`: Kiongozi wa chama cha Orange Democratic Movement ni nani?
`Positive document`: Orange Democratic Movement Katika uchaguzi wa rais Raila Odinga alitangazwa kuwa ameshindwa na rais Kibaki kwa kura 230,000. Lakini watazamaji wengi waliona kasoro katika hesabu ya kura na ODM ilidai kuwa Odinga ni mshindi halali. ODM ilifaulu vizuri upande wa viti vya bunge la Kenya. Ilipata karibu nusu ya wabunge wote yaani 99 kati ya 120 ikawa kubwa katika bunge baada ya uchaguzi wa Desemba 200 . . .

SPLARE | top features (doc rank = 5) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 8) Token	%
mentions of "Orange" or related terms and concepts	8.9	␣Rail	9.3
references to events or occurrences in the future	6.4	␣Orange	8.5
prominent political figures and their involvement in elections	6.0	␣Kenya	7.1
references to the abbreviation "OD" and variations of it, typically related to a specific . . .	5.8	Rail	6.3
terms associated with political events and discussions	5.4	OD	5.7
references to business strategies and company operations	5.1	␣OD	5.5
references to DMCA regulations and related legal terms	4.1	Orange	4.8
references to places in Kenya	4.0	␣movement	4.2
information related to notable historical figures and their relationships	3.6	␣leader	4.0
references to political candidates and their activities within the Democratic Party	3.3	␣orange	3.5

Figure 12: Retrieval example from `MIRACL/Bengali`

`Query (translated from Bengali)`: What is the name of the first band in Bangladesh?
`Positive document (translated from Bengali)`: Obscure (Bangla Band) — Obscure is one of the notable bands in the history of Bangladeshi band music. In the 1980s, Sayed Hasan Tipu took the initiative to establish this band. On March 15, 1985, Tipu founded Obscure in Khulna. During the 1980s, Obscure's first album was released from Sargam Studio. That first self-titled album, "Obscure Volume 1," released in 1986, earned a permanent place in the history of Bangla band music.

SPLARE | top features (doc rank = 3) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 20) Token	%
references to specific individuals and groups within a social or cultural context	6.2	␣Bangladesh	9.8
references to musical bands or groups	6.2	␣band	8.5
mentions of bands and musical groups	5.5	␣Bang	7.9
repeated or emphasized mentions of specific entities or concepts	4.8	Bang	5.8
references to iconic rock bands and their legacy	4.7	␣bang	5.5
occurrences of the country name "Bangladesh."	4.7	␣Band	5.3
references to musical bands and collaborations	4.5	band	4.8
elements related to music and musicians	3.4	␣bands	4.2
descriptors related to music and performance quality	3.3	bang	3.8
proper names and the mention of individuals in the text	3.2	-band	3.3

Figure 13: Retrieval example from `MIRACL/French`

`Query`: Qui est le mathématicien le plus célèbre au monde?
`Positive document`: Nira Chamberlain En 2017, il intervient dans l'atelier du New Scientist "Le monde mathématique". En 2018, il est nommé "mathématicien le plus intéressant du monde" par le "Big Internet Math Off" organisé par le site "Aperiodical". En 2019, il donne une conférence à la "Maxwell Society" sur "Les mathématiques qui peuvent arrêter une apocalypse de l'IA". Il fait des apparitions dans les médias brita . . .

SPLARE | top features (doc rank = 5) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (doc rank = 11) Token	%
concepts related to mathematics and quantitative analysis	9.7	␣mathematic	19.7
terms and phrases related to mathematics	9.0	␣Mathematic	15.3
elements related to academic papers and research acknowledgments	8.5	␣maths	10.3
references to the concept of "world" in various contexts	8.3	␣math	10.2
references to mathematical concepts and theorems	6.8	␣Math	8.1
discussions about artistic individuals or the concept of creativity	6.6	ian	7.9
terms and references related to mathematics and mathematicians	6.2	␣world	6.6
references to mathematical concepts and terms	6.1	␣monde	5.9
terms related to academic professionals and researchers across various fields	5.1	␣mathematical	4.4
references to notable individuals and their contributions or warnings in the field of arti . . .	5.0	ematik	3.9

Figure 14: Retrieval example from XTREME-UP: `Tamil` → `English`

`Query`: மனிதனால் சராசரியாக எவ்வளவு வெட்பநிலையை தாங்க முடியும்
`Translation`: On average, how much temperature can a human withstand?
`Positive document`: Cold and heat adaptations in humans The human body always works to remain in homeostasis. One form of homeostasis is thermoregulation. Body temperature varies in every individual, but the average internal temperature is 37.0 ℃ (98.6 ℉). Stress from extreme external temperature can cause the human body to shut down [...]

SPLARE | top features (document rank = 2) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (document rank = 73) Token	%
references to "human beings" and related concepts	12.3	␣human	29.3
terms related to temperature variations and environmental conditions	11.9	␣Human	21.5
terms related to fever and its physiological effects	10.3	␣average	20.7
terms related to biological concepts and interactions	9.6	␣humans	20.3
references to averages or average values in contexts related to statistics or metrics	9.2	␣withstand	3.8
specific guidelines and recommendations related to health and wellness	8.6	␣endurance	2.4
phrases related to summer and heat conditions	8.2	␣limit	2.1
specific temperature values and their measurements	6.8		
references to bodily systems and their components	6.7		
quantitative data related to spending and financial metrics	5.9		

Figure 15: from `CodeEditSearchRetrieval`

Query: `Add totally untested pools ;)`

Positive document:
```
---
+++
@@ -1,4 +1,6 @@
-import abc
+from multiprocessing import Pool as ProcessPool
+from multiprocessing.dummy import Pool as ThreadPool
+from multiprocessing import cpu_count
def do_flow(flow, result=None):
@@ -8,19 +10,41 @@
return result
+class PoolAPI(object):
+ def map(self, *args, **kw):
+ return self.pool.map(*args, **kw)
+
+
+class ThreadPool(PoolAPI):
+
+ de ...
```

SPLARE | **top features** (document rank = 10) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE** | **top tokens** (document rank = 3) Token	%
references to "pool" and related concepts in various contexts	32.4	␣pool	17.0
aspects of vacation experiences related to comfort and amenities	28.0	␣pools	15.1
references to specific events or actions occurring in a timeline or sequence	7.1	␣Pool	14.5
topics related to punk culture and its influence on music and community	6.9	pool	8.9
references to the term "stream" or its variations within contexts related to art and medi …	5.8	Pool	8.8
comments about changes and improvements, particularly in processes, products, or performa …	5.5	␣pooling	6.6
references to swimming pools and related recreational facilities	3.8	␣pool	3.9
references to programming tasks and contributions related to software development	3.7	␣patch	3.7
references to programming interfaces and database management	3.2	␣improvements	3.1
expressions of sports performance and competition	1.4	␣patches	2.4

Figure 16: from `CodeEditSearchRetrieval`

Query: `Make sure that the interests_register tables are created Nose tries to run the interests_register tests, but they will fail unless the interest_register app is added to INSTALLED_APPS, because its ta ...`

Positive document:
```
---
+++
@@ -8,7 +8,8 @@
'pombola.place_data',
'pombola.votematch',
'speeches',
- 'pombola.spinner' ) + \
+ 'pombola.spinner',
+ 'pombola.interests_register') + \
APPS_REQUIRED_BY_SPEECHES
# create the ENABLED_FEATURES hash that is used to toggle features on and off.
```

SPLARE | **top features** (document rank = 14) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE** | **top tokens** (document rank = 1) Token	%
keywords related to the concept of registration in various contexts	9.5	␣apps	5.7
phrases relating to various aspects of "interest" in different contexts	9.2	␣register	5.3
mentions of applications or software-related terminology	5.9	␣register	5.2
features related to software libraries and their installation	4.7	interest	5.1
mentions of interest in various contexts or subjects	4.0	register	5.0
instances of the word "create" and its variations, indicating a focus on creation and gen …	3.9	␣Interest	4.8
issues related to coding and technical errors in query parameters	3.9	␣interests	4.6
phrases related to registration and enrollment	3.0	apps	4.5
elements related to importing and structuring code within modules	2.8	␣interest	3.7
control keywords related to system configuration and management	2.8	␣interest	3.7

Figure 17: from `CodeEditSearchRetrieval`

Query: `Update variables names in exam tests`

Positive document:

```
---
+++
@@ -17,16 +17,16 @@
def test_create_biopsy_exam(self):
from biopsy.models import Biopsy
- specific_exam = create_specific_exam('Biopsy')
+ biopsy_exam = create_specific_exam('Biopsy')
- specific_exam | should | be_kind_of(Biopsy)
+ biopsy_exam | should | be_kind_of(Biopsy)
def test_create_necropsy_exam(self):
from necropsy.mod ...
```

SPLARE | top features (document rank = 19) Explanation (from Neuronpedia) Lin (2023)	%	**SPLADE | top tokens** (document rank = 1) Token	%
terms associated with analysis and examination in a specialized medical context	19.8	␣exam	12.4
instances of the word "exam."	17.2	␣tests	9.0
occurrences of the word "test" and its variations in various contexts	16.9	␣Exam	8.9
references to exams and testing processes	15.2	␣exams	7.4
references to testing and test cases in programming contexts	10.2	exam	6.4
references to unit testing and its associated concepts	9.6	␣test	6.2
references to notable figures or characters in a narrative context	3.2	Exam	5.3
technical terms and keywords related to programming and computer science	2.6	␣exam	4.6
references to institutions or organizations in a structured context	2.1	␣examination	4.3
numerical values and measurements	1.7	tests	4.0