Scientific knowledge injection and multilingual alignment for concept-driven retrieval with sentence embedding models

Nicolau Duran-Silva[1,2,*], Pablo Accuosto[1] and Horacio Saggion[2]

[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Accessing research and innovation information increasingly requires effective retrieval across languages, document types, and levels of textual granularity. In many research ecosystems, content is inherently multilingual and queries are short, concept-driven, and underspecified, posing challenges for traditional lexical retrieval methods, while performance of general-purpose dense retrieval is limited. In this work, we present an empirical evaluation of multilingual dense retrieval for scholarly documents in Catalan, Spanish, and English. We analyse the behaviour of general-purpose and domain-adapted embedding models across monolingual and cross-lingual settings, query types, and query lengths, and compare dense retrieval against strong sparse baselines. Using weakly supervised query–passage and triplet datasets derived from open research information, we show that domain-specific multilingual fine-tuning substantially improves retrieval effectiveness, semantic alignment, and embedding coherence. Our results highlight the importance of domain and multilingual adaptation for robust scholarly information access. These capabilities are particularly important for research mapping and scientometric analysis tools, where retrieval quality can directly influence downstream analytical modules such as topic and collaboration analysis, or research portfolio mapping.

Keywords

Scholarly Information Access, Dense Retrieval, Multilingual Semantic Search, Domain Adaptation

1. Introduction

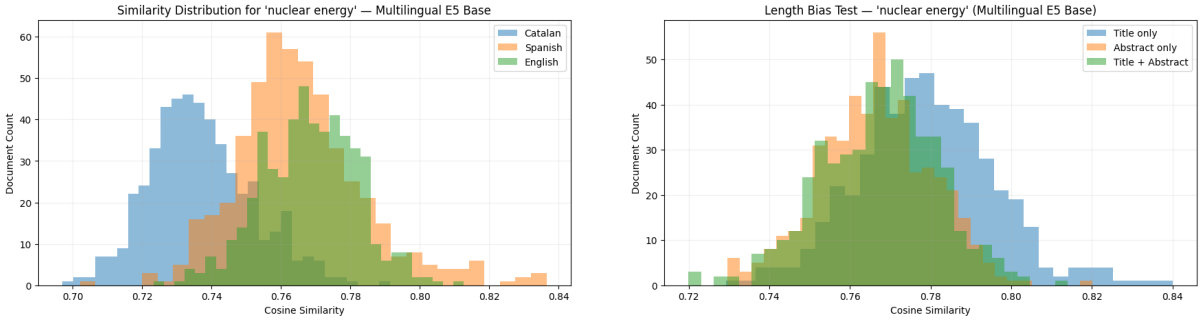
Scientific and technical information is increasingly available through open databases of research projects, scholarly publications, and patents [1, 2, 3], which contain an enormous quantity of textual information that details current challenges, proposed advancements, used technologies, and expected impact of the research and innovation process [4]. Given this situation, one could think that the growing amount of available information is very useful to foster new discoveries and advances in research. However, accessing and reading this large and growing amount of documents would be extremely time-consuming, and therefore, unfeasible for humans [5].

These documents form the basis of *research mapping platforms* [6, 7, 8] which allow researchers and policymakers to search, compare, and analyse of research and innovation activities and production across languages, institutions, and funding instruments. Search in scholarly and project repositories is therefore often multilingual and concept-driven. Research information systems aggregate outputs from different territories and communities with distinct dominant languages, and publicly funded research projects managed by national or regional funding agencies frequently provide titles and descriptions in local languages [9, 10, 6, 7, 11]. These challenges are particularly evident in publicly funded research data, where documents are distributed across local, national, and international repositories, and titles and descriptions may vary substantially in length, detail, and availability.

The aim of *research mapping platforms* often goes beyond traditional document ranking and *relevance*, because search results are also commonly used as input to analytical modules aimed at understanding

SCOLIA '26: Second International Workshop on Scholarly Information Access (SCOLIA), April 2, 2026, Delft, The Netherlands
*Corresponding author.
nicolau.duransilva@sirisacademic.com (N. Duran-Silva)
0000-0001-5170-4129 (N. Duran-Silva); 0000-0002-3493-6013 (P. Accuosto); 0000-0003-0016-7807 (H. Saggion)

scientific specialisation, organisational performance, or thematic trends across research portfolios. In this sense, retrieval can be interpred as a form of *classification*. Users frequently issue brief queries such as 'cancer', 'artificial intelligence' or 'blue economy', expecting to retrieve relevant documents that may be written in different languages and described using diverse and more specific scientific terminology. These fine-grained scientific concepts are the basis of scientific queries [12]. However, the major difficulty in scholarly information retrieval could be the knowledge behind the words which is expected to be known or understood [13], especially relevant for those queries that are not self-descriptive or only known by domain experts.



(a) Similarity distribution over the trilingual project dataset (Catalan, English, Spanish) to exemplify the language bias from the query.

(b) Similarity distribution restricted to English projects, comparing embeddings computed from title only, abstract only, and title+abstract to assess the effect of text length.

Figure 1: Cosine similarity distributions for the query "nuclear energy" with multilingual E5 [14] on a sample of 1.5k project descriptions in English, Spanish, and Catalan. Similarity scores are computed using cosine similarity between query and document embeddings produced by SentenceTransformers.

While traditional keyword-based information retrieval systems can handle lexical representations [15], they fail to capture the semantic relationships and contextual meaning that characterise modern scientific language—for example, they cannot recognise that 'oncology' and 'cancer research' refer to related concepts, or that a query in Spanish should match an English abstract on the same topic, or to process more complex queries like 'AI for energy transition'. Recent advances in multilingual large language models have enabled more natural and concept-oriented interaction with textual databases through semantic search [16]. Embedding-based retrieval [17, 18] represent search queries and documents in a shared semantic space, supporting retrieval beyond exact word matches, allowing proper semantic search. Sentence encoder models are increasingly used from scholarly RAG systems [19], to modern scientometrics topic modelling approaches [20, 21]. While language models are trained to represent context, *concept-based* queries (e.g. 'cancer') lack of sufficient context to produce informative dense representations.

However, semantic search systems often rely on similarity thresholds or ranking signals derived from embedding similarity, and general-purpose embedding models may introduce biases that affect scholarly retrieval effectiveness. Similarity scores in this analysis are computed using cosine similarity between query and document embeddings produced by SentenceTransformers [17]. As shown in Figure 1, despite being multilingual, the model displays cross-lingual differences in similarity distributions. In this case, documents written in the query language (English, in this case) tend to receive higher similarity scores, while texts in less-represented languages, such as Catalan, obtain lower scores. While retrieval metrics such as Recall@k and MRR depend on the relative ranking of documents rather than absolute similarity values, systematic shifts in similarity distributions may still influence ranking outcomes when relevant documents consistently receive lower similarity scores than competing candidates. In addition, similarity scores correlate with passage length, with shorter texts (e.g., titles) often ranked ahead of longer passages regardless of their semantic completeness given a concept-based query. This behaviour is something we have observed in the practical development of dense retrieval search systems with industry-standard models, as well as narrow similarity window discrimination for scientific documents

with general-purpose models. Although these behaviours are partly expected, this issue is particularly relevant in open scholarly knowledge graphs, where abstract availability has recently decreased due to publisher restrictions [22].

Our goal is to evaluate and improve dense retrieval models for multilingual documents by adapting models with:

- scientific domain knowledge,
- multilingual alignment across Catalan, Spanish, and English,
- ranking-oriented behaviour for search,
- fine-grained cosine separation to support classification and analytical tasks.

Our contribution is primarily empirical rather than architectural. Instead of proposing new training objectives or model architectures, we evaluate multilingual semantic search for scholarly publications and projects, analysing how general-purpose embedding models behave across languages, passage lengths, and similarity ranges, and how they can be adapted to the scientific multilingual domain. This study builds on our previous work on multilingual semantic retrieval and query segmentation for scientific information access, expanding both evaluation and experiments. The development of these models is in the context of building new text search capabilities for open research information platforms in Catalonia, with tools like RIS3MCAT [6][1], for this reason we focus on titles and abstracts from both research publications and funded projects in Catalan, Spanish, and English.

These capabilities are directly relevant to scientometric and research policy analysis. Research mapping platforms such as RIS3MCAT integrate search with analytical modules that support the exploration of research portfolios, collaboration networks, thematic specialisation, and funding impact. In such systems, retrieval acts as a filtering and classification step that determines which documents are included in downstream analytical workflows. Consequently, the quality of semantic retrieval directly affects the reliability of scientometric analyses derived from these platforms.

2. Related Work

Scholarly information access research has addressed the challenge raised by the rapid growth of scientific literature, by exploring how to address information needs of researchers [23], developing recommendation systems [24] and discovery tools [25], and using publication textual content and metadata [26]. Semantic similarity search is dominated by dense retrieval methods, which encode queries and documents into a shared embedding space and rank candidates by vector similarity [27, 28]. These approaches enable concept-level search beyond lexical overlap, but also face known limitations [29] in capturing rare terminology and fine-grained distinctions, issues especially relevant for scientific and multilingual contexts. However, several studies show that dense retrieval does not consistently outperform strong lexical baselines as BM25 [30]. However, highlight that fine-tuning a dense model on domain-specific data lead to improved performance, surpassing BM25 in most metrics. Dense retrieval models trained in one domain do not generalise properly in others [31], particularly general-purpose semantic representation models often fail to capture fine-grained scientific concepts [12].

A key challenge for dense retrieval of scholarly documents is the lack of annotated data for training and test. Creating supervised query-passage pairs for scientific literature is costly and generally requires domain experts [12]. To address this gap, prior studies have explored the challenge of generating training labels with unsupervised and weakly supervised approaches [32], including pseudolabelling strategies [31], automatic generation of negative examples [33], and query expansion with LLMs [12], Their significant improvements suggest that dense retrievers can be trained without manually labelled data. The challenge of weakly supervised dataset creation for scholarly document processing has been addressed due to the challenge and cost of generating those labels, which generally require domain experts [34, 35].

[1]Available at https://ris3mcat.gencat.cat/.

The sentence-transformers framework [17] provides a widely adopted pipeline for training dense retrievers, typically using Multiple Negatives Ranking Loss (MNRL) [36], where in-batch examples act as implicit negatives to efficiently learn discriminative representations. Recent advances refine contrastive objectives through better negative sampling, hard negative mining [37], cross-lingual pairs [38], and improved optimisation strategies [18, 39], all contributing to more robust vectorial representations. Hybrid retrieval architectures partially address the weaknesses of dense-only methods in handling exact matches and rare entities (e.g., uncommon organisation names, specialised technical terms, or newly coined concepts that appear infrequently in training corpora). Late-interaction models such as ColBERT [40], preserve token-level granularity while maintaining efficiency, and recent multi-vector approaches [41] further improve retrieval precision through fixed-dimensional encodings. Others [42] have explored extraction and indexing of relevant dimensions of scholarly abstracts like directions or challenges described. For deployment, vector indexes based on HNSW graphs [43] remain the standard for low-latency large-scale retrieval.

In the multilingual domain, several model families are particularly relevant. The multilingual E5 models [44] show strong cross-lingual transfer from large-scale retrieval corpora. Multilingual RoBERTa-based encoders trained in trilingual query relevance dataset (on 65k CA-ES-EN query-passage pairs) demonstrates effectiveness when trained with domain-appropriate data [45]. These models benefit substantially from domain-specific contrastive fine-tuning, which improves discrimination between closely related scientific concepts. While in scientific domain, SPECTER [46] leverages citation networks to specialise embeddings for scientific papers (though predominantly in English). However, recent work [12] compare E5 and Specter2 [18], finding E5 can achieve best results, better than BM25 or hybrid baselines.

Dense retrieval also plays a central role in retrieval-augmented generation (RAG) frameworks, improving factual accuracy for LLMs [47, 48]. However, adapting dense retrievers to specialised multilingual scientific domains remains challenging due to domain-specific terminology, code-switching, and limited non-English training data [49, 50]. Our approach follows the contrastive paradigm while introducing domain-specific multilingual pairs to strengthen semantic alignment across Catalan, Spanish, and English research texts.

3. Methods

This section describes the methodology used to evaluate and adapt dense retrieval models for multilingual scholarly search, with a focus on cross-lingual alignment, ranking behaviour, and sensitivity to passage length. Following prior studies [35, 32], we rely on weak supervision derived from latent and author-provided publication metadata and machine translation for training models. This setting reflects realistic constraints in multilingual scholarly information access, where large-scale expert annotation is not available. An alternative approach to multilingual retrieval would consist of translating all queries and documents into a pivot language such as English; however, in this work we focus on multilingual embeddings to avoid full-corpus translation and preserve original-language representations.

3.1. Retrieval Models

Base Models. We evaluate a set of multilingual or scientific-domain sentence encoder models:

- `Multilingual E5`[2] [44]: a multilingual text embedding encoder trainer on MS-MARCO dataset [51], a large-scale passage retrieval dataset derived from Bing search queries.
- `mRoBERTA_retrieval`[3]: a trilingual RoBERTa model pre-trained on CA, ES, and EN data.
- `distilRoBERTa`[4] [17]: a lightweight English sentence encoder used as general-purpose baseline.
- `SPECTER`[5] [46]: a scientific-domain encoder trained on English documents for citation similarity,

[2] huggingface.co/intfloat/multilingual-e5-base
[3] huggingface.co/langtech-innovation/mRoBERTA_retrieval
[4] huggingface.co/sentence-transformers/all-distilroberta-v1
[5] huggingface.co/sentence-transformers/allenai-specter

providing a strong baseline for semantic paper retrieval.

3.2. Datasets

We build several weakly supervised datasets from openly available scholarly collections of publications and projects to support training, evaluation and analysis.

Trilingual Research Project Corpus.
This is a dataset of 1.5K publicly funded research projects and is used to analyse retrieval behaviour across languages and document granularities. It consists of 500 English projects from the European Commission's CORDIS platform[6], 500 Catalan projects from RIS3CAT[6][7], and 500 Spanish projects from AEI[8] and CDTI[9]. Each record includes title and description, when available. We have manually annotated relevant documents according to 5 different concept-driven queries, using a pooling strategy (for each query, the top 30 candidate projects retrieved by keyword search, bm25, and dense model variants).

Query-Passage Dataset.[10]
Our primary training dataset comprises 76k query-text pairs, equally distributed across English, Catalan, and Spanish. We collect 30K scientific publications in English from several bibliographic databases[11], extracting their titles, abstracts and author keywords. Individual author keywords are treated as queries, while titles and abstracts serve as passages. To obtain multilingual supervision samples, all textual fields (author keywords, titles and abstracts) are automatically translated into Catalan and Spanish using machine translation system, using the Google Translate API. The original and translated texts are then aligned to construct both monolingual and cross-lingual query-passage pairs across the three languages. There are no repeated articles between languages. The dataset contains both monolingual and cross-lingual pairs. Approximately 90% of the examples correspond to `keyword→text` pairs (where the text may consist of the title, abstract, or title+abstract), reflecting the typical use of short concept queries in scholarly search, and the missing abstracts for some records. The remaining 10% consist of `title→abstract` pairs to preserve document-level semantic similarity during training, and for textual equivalent searches. While automatic translation may introduce some noise or semantic drift, keywords are typically short technical terms, which reduces the likelihood of substantial translation errors. Contrastive training has been shown to be robust to moderate noise in supervision signals. The overall data construction process is illustrated in Figure 2. This dataset can therefore be considered a weakly supervised resource, where author keywords act as implicit relevance signals. In a low-resource multilingual setting, this approach enable the generation of cross-lingual training pairs with a low investment of resources. The dataset is split into 80/10/10 partitions. Splitting is performed at pair level, ensuring that each query-passage pair appears in only one partition. Because queries correspond to author keywords representing scientific concepts, the same query term (e.g. "cancer") may occur across splits paired with different passages, while longer and more specific queries appear less frequently. This setup reflects realistic retrieval scenarios where common concepts may correspond to many different documents.

Classification Dataset.[12]
We additionally use the classification dataset `scidocs-mag` [46], translating one third to Catalan and one third Spanish, which are annotated with 19 scientific categories corresponding to Microsoft Academic Graph's Fields of Science at level 0. This is used for computing polarity score and optimal similarity threshold search.

[6] cordis.europa.eu/projects
[7] https://ris3mcat.gencat.cat/
[8] aei.gob.es/ayudas-concedidas/buscador-ayudas-concedidas
[9] cdti.es/datos-abiertos-creditos-subvenciones-y-lineas
[10] huggingface.co/datasets/nicolauduran45/multilingual_research_pairs
[11] huggingface.co/datasets/nicolauduran45/scidocs-keywords-exkeyliword
[12] huggingface.co/datasets/nicolauduran45/multilingual-research-classification

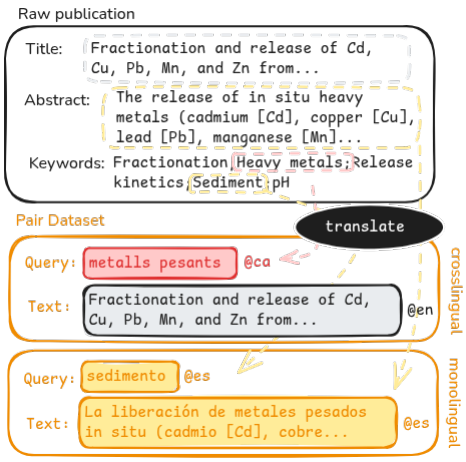


Figure 2: Examples of pairing mechanism for constructing the pair dataset.

3.3. Model Fine-tuning

In order to assess different fine-tuning strategies on our query–passage dataset, we experimented with different dataset configurations and loss functions. This allows us to analyse how training objectives influence multilingual alignment and ranking behaviour.

Loss Functions. We evaluate four complementary loss functions [17] commonly used in dense retrieval.

- *Multiple Negatives Ranking Loss (MNRL)* [52], which performs contrastive learning using in-batch negatives. For each query–passage positive pair, all other passages in the same mini-batch (batch size = 32) are treated as implicit negatives, without explicit negative sampling.
- *Contrastive Loss* [36], a pairwise objective that learns to separate relevant and non-relevant query–document pairs using a margin-based formulation. We adapt the dataset by pairing each query with its associated positive passage and sampling a single explicit negative passage per query. Negatives are selected randomly under the constraint that they do not share any annotated positive keywords with the query, ensuring semantically safe negative examples. Positive pairs are labelled with similarity 1 and negative pairs with 0.
- *Triplet Loss*, which explicitly enforces relative similarity constraints between queries, relevant passages, and hard negatives. For each query, we form triplets consisting of an anchor (query), a positive passage, and a sampled negative passage. Negative passages are selected as in Contrastive Loss. A cosine-distance margin of 0.5 is used.
- *Cosine Similarity Loss*, which optimises cosine similarity scores for query–document pairs. We use the same pairwise dataset construction as for Contrastive Loss.

Training Setup. Models are fine-tuned using `SentenceTransformer` framework [17]. Training is performed for three epochs with identical optimisation setting across models, and evaluated on held-out test data. Detailed hyperparameter settings are reported in Appendix B.

4. Evaluation and analysis

We explore model retrieval capacity, and analyse performance in monolingual and cross-lingual settings, explore the impact of query length, as well as model behaviour for lexical and semantic queries. We evaluate on 5 example queries performance of sparse and dense retrieval and explore how to choose the best similarity thresholds for dense retrieval.

4.1. Evaluation metrics

Models are evaluated on the held-out multilingual test split using the following metrics:

- **Top-k Recall** ($k \in \{1, 5, 10\}$): proportion of queries for which the paired passage appears among the top-k retrieved results.
- **Cosine@1**: average cosine similarity between positive query–passage pairs, measuring embedding alignment quality.
- **Mean Reciprocal Rank (MRR)**: evaluates the ranking quality by measuring the inverse rank of the first correct passage within the top-10 retrieved candidates.
- **Neighbourhood Polarity**: following [39] formulation, we compute the proportion of the top-k nearest neighbours in the embedding space that share the same class label (discipline) as the target document. Higher polarity indicates more coherent semantic neighbourhoods and stronger clustering of scientific topics.

4.2. Overall model performance

Table 1 reports retrieval performance of the base and fine-tuned embedding models on our multilingual *Query-Passage Dataset* test split. We report R@k, MRR, cosine accuracy, and neighbouring polarity. To ensure robust evaluation given the weakly supervised construction of the dataset, retrieval is performed over batches of 64 candidate documents. Because supervision signals are derived from author keywords, relevance annotations are incomplete, given that a document may be relevant to a query even if it is not paired in the dataset. This setting reduces false positives arising from semantically related but non-paired samples. Evaluating over the full corpus would introduce several false negatives and artificially penalise correct semantic matches. Restricting the candidate pool allows us to measure the capacity of the model to distinguish relevant passages from semantically related distractors while mitigating noise from incomplete supervision.

In addition, we also treat as valid positives any passages associated with additional author keywords from the same source document, reflecting the many-to-many nature of author keywords, where a single document may correspond to multiple conceptually related queries. The neighbouring polarity score derived from the classification dataset, measures whether the top-k neighbours (here, $k = 16$) share the same scientific field. This provides an external estimate of whether fine-tuned models preserve meaningful semantic structure across scientific domains.

Comparing loss functions, MNR loss consistently provides the largest gains, reflecting its explicit optimisation of relative similarity among in-batch candidates for retrieval. The results indicate that domain-specific and multilingual enrichment significantly improves both ranking and semantic organisation of the embedding space. While E5 achieves the strongest overall performance, all models benefit from fine-tuning with MNR loss, including encoders originally trained only on English data. In contrast, neighbourhood polarity improves to a similar extent across all loss functions, suggesting that most objectives encourage comparable levels of inter-document semantic cohesion, even when the improvements in retrieval are smaller.

In the following sections, we focus on the best-performing fine-tuned models, those with MNR loss, and analyse their behaviour in more detail.

4.2.1. Performance by Query Type

To better analyse retrieval behaviour, and to answer the question of how well dense retrieval preserve lexical retrieval capacities, we further distinguish between lexical and semantic query-passage matches. A pair is considered *lexical* when the query string appears verbatim in the passage (e.g., query "cancer" and passage containing "breast cancer"). Otherwise, it is labelled as *semantic* when retrieval success requires conceptual inference or paraphrasing (e.g., query "cancer" and passage mentioning "basal cell carcinoma"). This classification is language-agnostic: cross-lingual pairs are still considered lexical if translated forms match directly. Table 2 reports Recall@1 and Recall@10 across lexical and semantic

Model	R@1	R@5	R@10	MRR	Polarity (k=16)
– Base Models					
E5	.47	.70	.80	.58	.55
mRoBERTA	.34	.59	.71	.46	.63
DistilRoBERTa	.39	.58	.68	.49	.55
Specter	.27	.47	.59	.37	.52
– Fine-tuned with ContrastiveLoss					
E5	.49	.75	.84	.61	.69
mRoBERTA	.38	.70	.82	.52	.66
DistilRoBERTa	.38	.59	.70	.49	.56
Specter	.39	.69	.81	.53	.61
– Fine-tuned with CosineSimilarityLoss					
E5	.49	.74	.84	.61	.69
mRoBERTA	.38	.69	.82	.52	.67
DistilRoBERTa	.40	.61	.72	.50	.57
Specter	.38	.69	.81	.52	.61
– Fine-tuned with MNRLoss					
E5	.74	.90	.95	.81	.68
mRoBERTA	.65	.86	.92	.74	.66
DistilRoBERTa	.62	.81	.88	.71	.61
Specter	.61	.83	.89	.71	.62
– Fine-tuned with TripletLoss					
E5	.51	.78	.87	.63	.69
mRoBERTA	.40	.71	.83	.54	.66
DistilRoBERTa	.42	.66	.77	.53	.57
Specter	.38	.67	.79	.52	.59

Table 1
Multilingual semantic retrieval performance across base and fine-tuned embedding models, with 64 candidate documents per query. Metrics: Recall@k (R@1/5/10), Mean Reciprocal Rank (MRR), and neighbourhood polarity at k=16.

matches, comparing base models and MNRLoss finetuned. Base models show a pronounced gap between lexical and semantic performance, indicating a strong reliance on surface-level term overlap. Fine-tuning with MNRLoss substantially improves retrieval performance for both match types, with lexical and semantic recall doubling in most cases. These results suggest that training on a mixture of lexical and semantic query–passage pairs strengthens both exact-match sensitivity and deeper semantic generalisation.

Model	R@1	R@10
Base Models	**Lex.| Sem.**	**Lex.| Sem.**
E5	.48 | .34	.83 | .73
mRoBERTA	.29 | .27	.68 | .68
DistilRoBERTa	.37 | .29	.68 | .62
Specter	.21 | .19	.54 | .55
Fine-tuned Models		
E5	.79 | .64	.96 | .92
mRoBERTA	.69 | .54	.94 | .88
DistilRoBERTa	.66 | .50	.91 | .83
Specter	.63 | .51	.91 | .85

Table 2
Lexical vs. semantic Recall@1 and Recall@10 for base and fine-tuned models with MNRLoss.

Model	R@1	R@10
Base Models	**Mono.| Cross.**	**Mono.| Cross.**
E5	.54 | .27	.89 | .66
mRoBERTA	.31 | .25	.70 | .66
DistilRoBERTa	.40 | .25	.73 | .56
Specter	.25 | .15	.60 | .49
Fine-tuned Models		
E5	.75 | .67	.95 | .93
mRoBERTA	.65 | .57	.92 | .89
DistilRoBERTa	.64 | .51	.89 | .84
Specter	.62 | .52	.89 | .86

Table 3
Recall@1 and Recall@10 segmented by pair type (monolingual vs. cross-lingual), comparing base and fine-tuned models with MNRLoss. *M = monolingual | C = cross-lingual.*

4.2.2. Performance by Language Configuration

To analyse the impact of multilingualism on retrieval quality, we evaluate the models separately under *monolingual* and *cross-lingual* pairs. In the monolingual scenario, queries and passages are written in the same language, while the cross-lingual scenario contains pairs where the query and the target text are in different languages. This distinction allows us to measure both in-language semantic retrieval and the ability of models to align concepts across languages. Table 3 reports Recall@1 and Recall@10 under both conditions for all base and fine-tuned models. We observe how in fine-tuned models the gap between monolingual and cross-lingual performances are reduced considerably.

4.2.3. Monolingual Performance by Language and Query Length

We further analyse monolingual retrieval performance by grouping test pairs according to the language of the target passage. This analysis examines how models handle scientific text in each language independently, isolating retrieval accuracy from cross-lingual alignment effects. Table 4 reports Recall@1 and Recall@10 for all models across the three languages. We observe the English as the dominant language, likely due to a major representation in training datam. However, this gap is reduced after fine-tuning, by most models Catalan is the worst, possibly due to representation and resource availability. Finally, we analyse in Table 5 retrieval performance in function of query length. Queries are grouped into three categories: short (single token), medium (2-3 tokens), and long (4 tokens). This captures the effect of some of the challenges of concept-driven keyword searches, comparing from no context to more descriptive queries. Across all models, fine-tuning yields substantial improvements for all lengths of queries, indicating enhanced robustness to limited contextual information.

Model	R@1	R@10
Base Models	**CA| EN| ES**	**CA| EN| ES**
E5	.45 | .62 | .54	.86 | .90 | .89
mRoBERTA	.26 | .35 | .32	.66 | .75 | .69
DistilRoBERTa	.31 | .59 | .29	.66 | .84 | .68
Specter	.18 | .34 | .21	.54 | .70 | .55
Fine-tuned Models		
E5	.70 | .79 | .75	.94 | .96 | .93
mRoBERTA	.60 | .67 | .67	.90 | .94 | .92
DistilRoBERTa	.57 | .73 | .61	.85 | .94 | .86
Specter	.55 | .71 | .59	.86 | .94 | .88

Table 4
Monolingual Recall@1 and Recall@10 across CA/EN/ES pairs, comparing base and fine-tuned models with MNRLoss.

Model	R@1			R@10		
Base Models	**Short**	**Med.**	**Long**	**Short**	**Med.**	**Long**
E5	.43	.58	.70	.82	.91	.97
mRoBERTA	.21	.35	.43	.58	.75	.85
DistilRoBERTa	.32	.43	.55	.67	.75	.83
Specter	.14	.29	.39	.47	.66	.75
Fine-tuned Models						
E5	.66	.78	.87	.90	.97	.98
mRoBERTA	.53	.69	.76	.87	.94	.97
DistilRoBERTa	.53	.68	.76	.83	.91	.95
Specter	.49	.68	.75	.83	.92	.95

Table 5
Monolingual retrieval performance by query length. Recall@1 and Recall@10 are reported for short (1 token), medium (2–3 tokens), long (4 tokens) queries. Comparing base and fine-tuned models with MNRLoss.

4.3. Comparing sparse and dense retrieval

To compare dense retrieval with well-established sparse methods, we suggest and conduct a small-scale analysis using exact keyword matching and BM25 as baselines. We evaluate five representative, concept-driven queries we have annotated over the trilingual research project corpus, spanning well-established scientific topics, emerging policy-oriented concepts, and semantically complex queries that are not always lexically explicit in project descriptions. Table 6 reports Precision@10 across retrieval methods. While BM25 provides a strong and stable baseline, particularly for topics with consistent terminology, base embedding models do not consistently outperform it. In contrast, fine-tuned embedding models, especially E5, achieve higher precision across all queries, with the largest gains observed for concept-driven queries of different natures. This is a small-scale analysis, but it would be interesting to analyse more deeply going forward, also as query routing strategies.

Method	'Nuclear Energy'	'Breast Cancer Research'	'Sustainable Food'	'Blue Economy'	'AI for Energy Transition'
Exact Keyword Match	.00	.00	.00	.20	.00
BM25	.20	.50	**.90**	.20	.20
E5 (base)	.00	.30	.30	.20	.20
mRoBERTa (base)	.10	.10	**.90**	.00	**.60**
DistilRoBERTa (base)	.00	.50	.80	**.40**	.20
Specter (base)	.10	.50	.50	.20	.00
E5 (ft)	**.50**	**.80**	**.90**	**.40**	**.60**
mRoBERTa (ft)	.10	.30	**.90**	.30	.50
DistilRoBERTa (ft)	.20	.70	.70	.10	.00
Specter (ft)	.40	.60	.80	.20	.00

Table 6
Precision@10 for different retrieval methods across five queries on a collection of 1.5k research projects.

4.4. Selecting Similarity Thresholds for Retrieval and Classification

A key challenge in dense retrieval is determining an appropriate relevance threshold on cosine similarity, particularly when retrieval outputs are used for analytical tasks. Unlike ranking-based evaluation, these applications require a binary decision on document relevance, making threshold selection both critical and model-dependent. To address this question, we leverage the classification dataset to estimate cosine similarity thresholds that maximise the F1 score on the test set. We report, in Table 7, average optimal `threshold` and corresponding F1 score across 19 subject categories, providing practical orientation and empirical guidelines for selecting similarity thresholds under different embedding models.

Model	Base		Fine-tuned	
	Threshold	F1	*Threshold*	F1
E5	.79	$.26 \pm .10$	.27	$.60 \pm .17$
mRoBERTA	.70	$.40 \pm .15$	.34	$.55 \pm .18$
DistilRoBERTa	.21	$.40 \pm .11$	.31	$.53 \pm .16$
Specter	.71	$.24 \pm .13$	.28	$.51 \pm .14$

Table 7
Average optimal similarity *threshold* and F1 score (mean $\pm$ standard deviation) across 19 categories of the test set of the classification dataset for base and fine-tuned embedding models.

To present the effect of threshold selection, we present an example in Figure 3, which visualises cosine similarity distributions with True/False samples for a representative query (*Environmental science*) under the base and fine-tuned E5 models. The histogram highlights how fine-tuning increases the separation between relevant and non-relevant documents, leading to higher F1 score.

5. Discussion

Across all experiments, fine-tuning consistently improves retrieval quality for every model and evaluation setting. Gains are especially pronounced in cross-lingual retrieval, where base encoders struggle to align Catalan, Spanish, and English scientific content. Models show 20–30 point improvements in Recall@1 after contrastive fine-tuning, confirming that domain-specific multilingual adaptation is important for multilingual scientific search.

Improvements extend across match types, while lexical queries are naturally easier, fine-tuning also boosts semantic retrieval capacity, demonstrating that the models learn to generalise beyond surface forms. Importantly, dense models do not lose lexical retrieval capacity, fine-tuning strengthens both lexical and semantic abilities. Even weaker base encoders become competitive multilingual

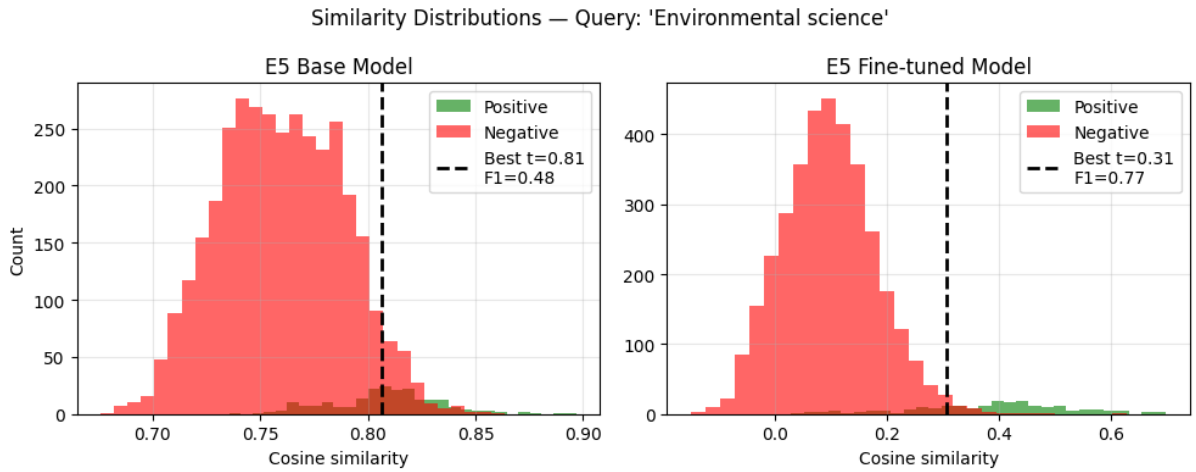


Figure 3: Cosine similarity score with base and fine-tuned E5 distributions from the classification dataset. Vertical dashed lines indicate the similarity thresholds that maximise the F1 score on the test set. Query = "Environmental science", Base Model (Best threshold = .807, F1 = .479), Ft Model (Best threshold = .306, F1 = .774)

retrievers after adaptation. Monolingual performance by language shows clear asymmetries that reflect underlying resource availability. English remains the easiest setting, with the highest scores even before adaptation. Catalan and Spanish lag behind in base models, particularly Catalan, which suffers from limited representation in openly available corpora. After fine-tuning, however, these gaps narrow substantially: Catalan gains the largest relative improvements, and Spanish reaches parity with English in some models.

These results show that the combination of multilingual contrastive learning and modest domain-specific supervision yields robust multilingual and cross-lingual semantic search capabilities—crucial for accessing R&I information in ecosystems where English, Spanish, and Catalan coexist. These models are able to improve upon strong sparse retrieval baselines. Finally, our analysis highlights that effective scholarly retrieval requires not only strong ranking performance but also interpretable similarity scores. The use of classification dataset threshold identification provides guides for bridging retrieval and analytical applications. Because from a scientometric perspective, improving retrieval quality is critical for research intelligence platforms rely on search results as input for analytical modules and facets that compute indicators such as thematic specialization or collaboration networks. Reliable multilingual retrieval therefore supports more accurate mapping and monitoring of research ecosystems.

6. Conclusion

In this work, we examine the performance of multilingual embedding models for accessing scientific and innovation data in a trilingual setting characteristic of many R&I information systems. Our results demonstrate that lightweight and domain-adapted models, including Catalan-centric variants, can effectively adapt to domain-specific data. Beyond findings, we contribute new multilingual datasets, model checkpoints, and evaluation resources designed to support future research on cross-lingual scientific information access. Taken together, our work underscores the importance of domain-specific adaptation and robust multilingual alignment for enabling reliable and scalable access to open research information. Beyond improving document retrieval, these capabilities are also relevant for scientometrics and research analysis, where semantic search systems are used to identify and analyse resarch portfolios and collaboration patterns.

Declaration on Generative AI

During the preparation of this work, the authors used the following generative AI tools and services: ChatGPT, Claude, DeepL, and LanguageTool. These tools were used exclusively to support writing-related tasks, including grammar and spelling checking, paraphrasing and sentence rephrasing, and general proofreading of the manuscript. In addition, generative AI tools were used for assistance in code development, documentation, and testing during the preparation of experimental scripts. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (grant reference 2022/DI /00017).

We thank the anonymous reviewers for their constructive feedback and suggestions, which improved the clarity and quality of this work.

References

[1] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, P. Príncipe, The openaire research graph data model, 2019. URL: https://api.semanticscholar.org/CorpusID:182277225.

[2] J. Priem, H. A. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, ArXiv abs/2205.01833 (2022). URL: https://api.semanticscholar.org/CorpusID:248512771.

[3] Z. Lin, Y. Yin, L. Liu, D. Wang, Sciscinet: A large-scale open data lake for the science of science research, Scientific Data 10 (2023) 315. doi:10.1038/s41597-023-02198-9.

[4] E. Fuster, F. A. Massucci, M. Matusiak, Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses, in: Quantitative Methods for Place-Based Innovation Policy, Edward Elgar Publishing, 2020, pp. 195–234. doi:10.4337/9781789905519.00014.

[5] I. Frommholz, P. Mayr, G. Cabanac, S. Verberne, C. K. Kreutz, The first workshop on scholarly information access (scolia), in: European Conference on Information Retrieval, Springer, 2025, pp. 326–331. doi:10.1007/978-3-031-88720-8_50.

[6] E. Fuster, T. Fernández, H. Carretero, N. Duran-Silva, R. Guixé, J. Pujol, B. Rondelli, G. Rull, M. Cortijo, M. Romagosa, Towards building a monitoring platform for a challenge-oriented smart specialisation with ris3-mcat, arXiv preprint arXiv:2401.10900 (2023). doi:10.48550/arXiv.2401.10900.

[7] ART-ER, SIRIS Academic, Monitoring Platform: Methodology Document – Smart Specialization Strategy 2021–2027, Technical Report, Emilia-Romagna Region, 2024. URL: https://monitoraggios3.art-er.it/documents/metodologia/S3%20Monitoring%20Methodology%20document.pdf, platform release updated as of November 5, 2024.

[8] D. Chaves, Product: The lens–patent and scholarly search analysis, Journal of the Canadian Health Libraries Association (JCHLA) 46 (2025).

[9] P. Baruch, Open access developments in france: the hal open archives system, Learned Publishing 20 (2007) 267–282. doi:10.1087/095315107X239636.

[10] S. M. d. Santos, G. Fraumann, S. Belli, R. Mugnaini, The relationship between the publication language and its impact on public and collective health (2020). doi:https://doi.org/10.1590/SciELOPreprints.1549.

[11] A. L. Packer, Multilingualism in scientific literature communicated by journals from the scielo brazil collection, European Review 32 (2024) S124–S144. doi:10.1017/S1062798724000103.

[12] Y. Zhang, R. Yang, S. Jiao, S. Kang, J. Han, Scientific paper retrieval with llm-guided semantic-based ranking, arXiv preprint arXiv:2505.21815 (2025). doi:10.48550/arXiv.2505.21815.

[13] C. Friedman, P. Kra, A. Rzhetsky, Two biomedical sublanguages: a description based on the theories of zellig harris, Journal of biomedical informatics 35 (2002) 222–235. doi:10.1016/S1532-0464(03)00012-1.

[14] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024). doi:10.48550/arXiv.2402.05672.

[15] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, volume 4, Now Publishers Inc, 2009. doi:10.1561/1500000019.

[16] A. Biswal, L. Patel, S. Jha, A. Kamsetty, S. Liu, J. E. Gonzalez, C. Guestrin, M. Zaharia, Text2sql is not enough: Unifying ai and databases with tag, arXiv preprint arXiv:2408.14717 (2024). doi:10.48550/arXiv.2408.14717.

[17] N. Reimers, I. 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, 2019. URL: https://arxiv.org/abs/1908.10084.

[18] A. Singh, M. D'Arcy, A. Cohan, D. Downey, S. Feldman, Scirepeval: A multi-format benchmark for scientific document representations, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5548–5566. doi:10.18653/v1/2023.emnlp-main.338.

[19] M. D. Skarlinski, S. Cox, J. M. Laurent, J. D. Braza, M. Hinks, M. J. Hammerling, M. Ponnapati, S. G. Rodriques, A. D. White, Language agents achieve superhuman synthesis of scientific knowledge, arXiv preprint arXiv:2409.13740 (2024). doi:10.48550/arXiv.2409.13740.

[20] A. Glazkova, Identifying topics of scientific articles with bert-based approaches and topic modeling, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2021, pp. 98–105. doi:10.1007/978-3-030-75015-2_10.

[21] N. Bovenzi, N. Duran-Silva, F. A. Massucci, F. Multari, J. Pujol-Llatse, Mapping sti ecosystems via open data: overcoming the limitations of conflicting taxonomies. a case study for climate change research in denmark, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2022, pp. 495–499. doi:10.1007/978-3-031-16802-4_52.

[22] B. Kramer, More open abstracts? comparing abstract coverage in crossref and openalex, 2024. URL: https://doi.org/10.5281/zenodo.11580550. doi:10.5281/zenodo.11580550.

[23] I. Frommholz, P. Mayr, G. Cabanac, S. Verberne, Bibliometric-enhanced information retrieval: 14th international bir workshop (bir 2024), in: European Conference on Information Retrieval, Springer, 2024, pp. 442–446. doi:10.1007/978-3-031-56069-9_61.

[24] S.-Y. Yang, C.-L. Hsu, S.-H. Lu, Developing an ontology-supported information recommending system for scholars, in: 2009 Joint Conferences on Pervasive Computing (JCPC), 2009, pp. 223–228. doi:10.1109/JCPC.2009.5420185.

[25] S. Volkova, P. Bautista, A. Hiriyanna, G. Ganberg, I. Erickson, Z. Klinefelter, N. Abele, H.-T. Kao, G. Engberson, Cross-disciplinary knowledge retrieval and synthesis: A compound ai architecture for scientific discovery, arXiv preprint arXiv:2511.18298 (2025). doi:10.48550/arXiv.2511.18298.

[26] T. Strohman, W. B. Croft, D. Jensen, Recommending citations for academic papers, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 705–706. doi:10.1145/1277741.1277868.

[27] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: EMNLP (1), 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[28] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, arXiv preprint arXiv:2112.09118 (2021). doi:10.48550/arXiv.2112.09118.

[29] O. Weller, M. Boratko, I. Naim, J. Lee, On the theoretical limitations of embedding-based retrieval, 2025. URL: https://arxiv.org/abs/2508.21038.

[30] L. Mori, C. Sousa de Oliveira, Y. Yih, M. Ventresca, Assessing the performance gap between lexical and semantic models for information retrieval with formulaic legal language, in: Proceedings of the Twentieth International Conference on Artificial Intelligence and Law, ICAIL '25, Association for Computing Machinery, New York, NY, USA, 2026, p. 114–128. URL: https://doi.org/10.1145/3769126.3769205. doi:10.1145/3769126.3769205.

[31] N. Thakur, N. Reimers, J. Lin, Injecting domain adaptation with learning-to-hash for effective and efficient zero-shot dense retrieval, arXiv preprint arXiv:2205.11498 (2022). doi:10.48550/arXiv.2205.11498.

[32] D. Li, V. Yadav, Z. Afzal, G. Tsatsaronis, Unsupervised dense retrieval for scientific articles, in: Y. Li, A. Lazaridou (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Abu Dhabi, UAE, 2022, pp. 313–321. URL: https://aclanthology.org/2022.emnlp-industry.32/. doi:10.18653/v1/2022.emnlp-industry.32.

[33] A. Sinha, P. S, R. Balaji, N. Bhatt, Bica: Effective biomedical dense retrieval with citation-aware hard negatives (2025). doi:10.48550/arXiv.2511.08029.

[34] A. Yakimovich, A. Beaugnon, Y. Huang, E. Ozkirimli, Labels in a haystack: Approaches beyond supervised learning in biomedical applications, Patterns 2 (2021). doi:10.1016/j.patter.2021.100383.

[35] M. Pàmies, J. Llop, F. Multari, N. Duran-Silva, C. Parra-Rojas, A. Gonzalez-Agirre, F. A. Massucci, M. Villegas, A weakly supervised textual entailment approach to zero-shot text classification, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 286–296. doi:10.18653/v1/2023.eacl-main.22.

[36] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[37] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, A. Overwijk, Approximate nearest neighbor negative contrastive learning for dense text retrieval, arXiv preprint arXiv:2007.00808 (2020). doi:10.48550/arXiv.2007.00808.

[38] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891. doi:10.18653/v1/2022.acl-long.62.

[39] T. E. Jørgensen, J. Breitung, Margins in contrastive learning: Evaluating multi-task retrieval for sentence embeddings, in: R. Johansson, S. Stymne (Eds.), Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), University of Tartu Library, Tallinn, Estonia, 2025, pp. 269–278. URL: https://aclanthology.org/2025.nodalida-1.28/.

[40] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48. doi:10.1145/3397271.3401075.

[41] L. Dhulipala, M. Hadian, R. Jayaram, J. Lee, V. Mirrokni, Muvera: multi-vector retrieval via fixed dimensional encodings, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Curran Associates Inc., Red Hook, NY, USA, 2024. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/b71cfefae46909178603b5bc6c11d3ae-Paper-Conference.pdf.

[42] D. Lahav, J. S. Falcon, B. Kuehl, S. Johnson, S. Parasa, N. Shomron, D. H. Chau, D. Yang, E. Horvitz, D. S. Weld, et al., A search engine for discovery of scientific challenges and directions, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11982–11990. doi:10.1609/aaai.v36i11.21456.

[43] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 824–836. URL: https://doi.org/10.1109/TPAMI.2018.2889473. doi:10.1109/TPAMI.2018.2889473.

[44] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A

technical report, 2024. URL: https://arxiv.org/abs/2402.05672.
[45] C. Rodriguez-Penagos, C. Armentano-Oller, M. Villegas, M. Melero, A. Gonzalez, O. d. G. Bonet, C. C. Pio, The catalan language club, arXiv preprint arXiv:2112.01894 (2021). doi:10.48550/arXiv.2112.01894.
[46] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2270–2282. doi:10.18653/v1/2020.acl-main.207.
[47] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
[48] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al., Improving language models by retrieving from trillions of tokens, in: International conference on machine learning, PMLR, 2022, pp. 2206–2240. URL: https://proceedings.mlr.press/v162/borgeaud22a.html.
[49] X. Zhang, X. Ma, P. Shi, J. Lin, Mr. TyDi: A multi-lingual benchmark for dense retrieval, in: D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, G. G. Sahin (Eds.), Proceedings of the 1st Workshop on Multilingual Representation Learning, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 127–137. URL: https://aclanthology.org/2021.mrl-1.12/. doi:10.18653/v1/2021.mrl-1.12.
[50] R. Litschko, I. Vulić, G. Glavaš, Parameter-efficient neural reranking for cross-lingual and multilingual retrieval, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 1071–1082. URL: https://aclanthology.org/2022.coling-1.90/.
[51] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al., Ms-marco: A human generated machine reading comprehension dataset, arXiv preprint arXiv:1611.09268 (2016). doi:10.48550/arXiv.1611.09268.
[52] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, R. Kurzweil, Efficient natural language response suggestion for smart reply, arXiv preprint arXiv:1705.00652 (2017). doi:10.48550/arXiv.1705.00652.

A. Online Resources

The datasets and models are available at:

- GitHub,
- Datasets & Models.

B. Fine-tuning Hyperparameters

We provide experimental details of our baseline fine-tuning approaches of sentence encoder models for content retrieval. Training was run (using 1x 24 GB GPU) for all models with hyperparameter defined in Table 8 .

Parameter	Setting
Loss function	4 different losses
Epochs	3
Batch size	32 per device
Learning rate	2e-5 (with 0.1 warm-up ratio)
Selection criterion	Best model selected based on R@1

Table 8
Fine-tuning configuration for multilingual document retrieval models.