Multilingual Generative Retrieval via Cross-lingual Semantic Compression

Yuxin Huang[1,2], **Simeng Wu**[1,2], **Ran Song**[1,2*], **Yan Xiang**[1,2],
Yantuan Xian[1,2], **Shengxiang Gao**[1,2], **Zhengtao Yu**[1,2]
[1]Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
{huangyuxin2004, simengggwu, song_ransr}@163.com, sharonxiang@126.com,
xianyt@kust.edu.cn, {gaoshengxiang.yn, ztyu}@hotmail.com

Abstract

Generative Information Retrieval is an emerging retrieval paradigm that exhibits remarkable performance in monolingual scenarios. However, applying these methods to multilingual retrieval still encounters two primary challenges, cross-lingual identifier misalignment and identifier inflation. To address these limitations, we propose Multilingual Generative Retrieval via Cross-lingual Semantic Compression (MGR-CSC), a novel framework that unifies semantically equivalent multilingual keywords into shared atoms to align semantics and compresses the identifier space, and we propose a dynamic multi-step constrained decoding strategy during retrieval. MGR-CSC improves cross-lingual alignment by assigning consistent identifiers and enhances decoding efficiency by reducing redundancy. Experiments demonstrate that MGR-CSC achieves outstanding retrieval accuracy, improving by 6.83% on mMarco100k and 4.77% on mNQ320k, while reducing document identifiers length by 74.51% and 78.2%, respectively. We publicly release our dataset and code at https://github.com/simengggg/MGR-CSC

1 Introduction

Multilingual Information Retrieval (MIR) serves as a critical component in natural language processing, particularly in applications such as cross-border e-commerce (Li et al., 2020) and cross-lingual search systems (Xu et al., 2021). The core need lies in developing models that can effectively process multilingual queries and retrieve relevant documents across different languages (Zhang et al., 2019; Dwivedi and Chandra, 2016). Traditional translation-based approaches compromise retrieval quality due to error propagation in machine translation pipelines (Chandra and Dwivedi, 2017). Recently multilingual pre-trained language models (PLMs) (Xue et al., 2020; Conneau et al., 2019) demonstrate improved performance by encoding cross-lingual content into joint semantic spaces (Yarmohammadi et al., 2019). However, precise cross-lingual alignment and end-to-end semantic matching remain significant challenges for pervious methods.

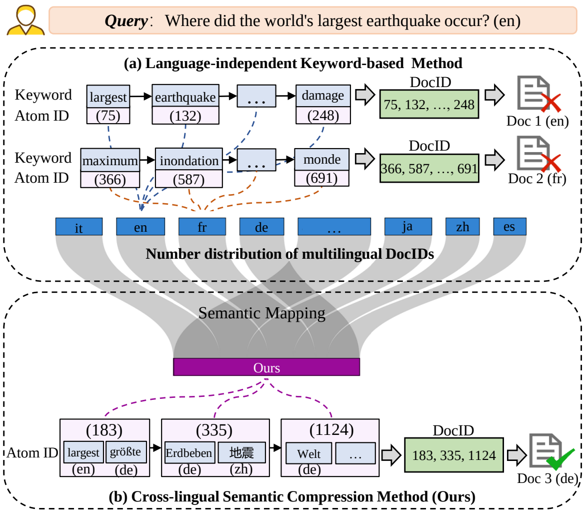


Figure 1: (a) illustrates the language-independent keyword approach, where the model biases toward query-language DocIDs during decoding. In contrast, (b) demonstrates MGR-CSC successfully retrieves target documents via semantic clustering, enabling more reliable identification.

Generative Information Retrieval (GIR) offers a paradigm shift by leveraging the model's parametric memory to store documents and directly generates document identifiers (DocIDs) (Tay et al., 2022; Zhuang et al., 2022; Sun et al., 2023). This approach leverages generative PLMs to learn and encode direct associations between documents and their unique DocIDs. Unlike previous paradigms, GIR offers an alternative end-to-end framework by utilizing generative PLMs to directly map queries to relevant DocIDs. This generative approach inherently addresses the mentioned challenges by

[*] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10855–10866
November 4-9, 2025 ©2025 Association for Computational Linguistics

learning direct associations between multilingual queries and DocIDs.

However, the application of GIR to multilingual scenarios confronts two fundamental limitations: (i) **Cross-lingual Identifier Misalignment.** Existing DocIDs are constructed using language-independent encoding schemes, creating isolated semantic mapping spaces for each language. As shown in Figure 1 (a), queries in a specific language lead the model to preferentially generate DocIDs for documents in the same language, such as English and French. This phenomenon is a direct result of cross-lingual identifier misalignment, which consequently hinders the transfer of multilingual knowledge through shared latent document-level representations. (ii) **Multilingual Identifier Inflation.** The number of existing DocIDs increases manifold as the number of languages grows in multilingual GIR. As shown in Figure 1 (a), keyword-based methods assign different atom IDs to semantically equivalent keywords across languages. For instance, the English word *largest* might be assigned atom ID *75*, while its French counterpart *maximum* is assigned atom ID *366*, despite their semantic equivalence. The combinatorial expansion of unique DocIDs in multilingual documents intensifies this challenge, posing significant hurdles for memory efficiency and runtime performance in autoregressive decoding architectures. Therefore, a semantically consistent DocID is essential for cross-lingual alignment and efficient decoding.

In this paper, we present **M**ultilingual **G**enerative **R**etrieval via **C**ross-lingual **S**emantic **C**ompression (MGR-CSC), a novel framework that unifies multilingual keyword semantics into shared atom IDs, assigns DocIDs to documents, and employs multi-step constrained decoding. A high-level overview of MGR-CSC is shown in Figure 1 (b). MGR-CSC performs cross-lingual semantic compression by mapping semantically equivalent multilingual keywords to shared atom IDs. For example, both the English word *largest* and its German counterpart *größte* are mapped to atom ID *183*. In addition, it applies dynamic decoding constraints to guide generation. Specifically, MGR-CSC contain three key parts. First, we extract explicit keywords from each multilingual document, and the document is represented by a set of multiple keywords. Secondly, these keywords are projected into a shared latent space using unsupervised clustering. Within this latent space, semantically equivalent expressions are assigned the same atom ID, which effectively compresses the multilingual identifier space. Finally, we introduce a multi-step dynamic constraint decoding strategy. The initial decoding step leverages the global frequency distribution of atom IDs to guide selection. And a refinement step narrows the selection space based on constraints between atomic IDs from preceding steps. Comprehensive experiments on multiple benchmark datasets show that MGR-CSC achieves outstanding performance, surpassing existing multilingual generative retrieval approaches by 6.83% on mMarco100k and 4.77% on mNQ320k. Furthermore, it substantially reduces the number of DocID tokens by 74.51% on mMarco100k and 78.2% on mNQ320k.

The contributions of this paper are as follows:

- We propose DocID construction approach for multilingual documents in MGR-CSC, enabling alignment and compact representation across languages.
- We propose a dynamic constrained multi-step decoding framework in MGR-CSC to reduce decoding complexity.
- Our experiments on multilingual benchmarks demonstrate the method's effectiveness and generalization in cross-lingual retrieval.

2 Related Work

2.1 Multilingual Information Retrieval

Multilingual information retrieval (MIR) seeks to semantically align queries and documents across languages, enabling cross-lingual access. Early MIR relied on translation, such as queries (Elayeb et al., 2018; Chandra and Dwivedi, 2017), documents (Yarmohammadi et al., 2019), or both (Dwivedi and Chandra, 2016) to balance efficiency and accuracy in practical systems. The advent of multilingual pretrained language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019) shifted attention to vector-based retrieval. In this paradigm, queries and documents are embedded into a shared space for similarity comparison (Yu and Allan, 2020; Zhang et al., 2022), thus eliminating external translation (Oard et al., 2008) and improving understanding in low-resource languages. Nonetheless, these approaches are limited by a fixed encode–match–rank pipeline lacking end-to-end optimization (Tay et al., 2022; Sun et al., 2023) and by contrastive learning's dependence on scarce parallel data (Karpukhin et al., 2020).

2.2 Generative Information Retrieval

Generative approaches are applied in fields such as information retrieval (Tay et al., 2022; Sun et al., 2023) and knowledge graphs(Song et al., 2024a; Abu-Rasheed et al., 2024). By capitalizing on the memorization capabilities of pre-trained language models(Song et al., 2024b), GIR directly generates DocIDs during inference, facilitating an end-to-end retrieval process. Current DocID representations primarily fall into two categories: atomic DocIDs and string DocIDs. For atomic DocIDs, Tay et al. (Tay et al., 2022) proposed constructing DocIDs using randomly assigned identifiers or cluster-based embedding layers. In contrast, string DocIDs employ semantically meaningful strings, such as document title (Tang et al., 2023) or keywords, such as TSGen (Zhang et al., 2024), Novo (Wang et al., 2023). Although, current GIR are primarily designed for monolingual settings, making effective adaptation to multilingual contexts difficult to achieve. To bridge this gap, we propose a semantic compression approach unifying cross-lingual lexical representations into a shared semantic space for multilingual GIR.

3 Methodology

In this section, we elaborate on our proposed method, Multilingual Generative Retrieval via Cross-lingual Semantic Compression, termed **MGR-CSC**. The core idea is to assign an unique DocID to each document by leveraging semantically similar key information across documents in different languages. This shared representation enables effective cross-lingual alignment and facilitates end-to-end semantic matching within a unified retrieval framework.

As Figure 2 illustrates, MGR-CSC consists of three components: (1) extracting distinctive keywords from multilingual documents, (2) clustering multiple keywords into one semantic atom, and assigning each document a unique DocID as a sequence of atoms, (3) during retrieval, the model generates the DocID atom-by-atom under dynamic decoding constraints.

3.1 Multilingual Keyword Extraction

To capture the essential semantics of multilingual documents, we extract a fixed set of m keywords from each document using a prompt-based Large Language Model (LLM). These keywords serve as a compact representation of the multilingual document's content.

Formally, given a document d_i, we extract a fixed number m of keywords as follows:

$$K_i = \{k_i^1, k_i^2, \dots, k_i^m\} = \text{LLM}(d_i), \qquad (1)$$

where K_i is treated as a semantically compact representation of the document d_i. And k_i^m denotes the m-th keyword of document i. We utilize a standardized extraction process across all language documents in order to ensure the consistency of semantic representations.

3.2 Semantic Atom Construction and DocID Assignment

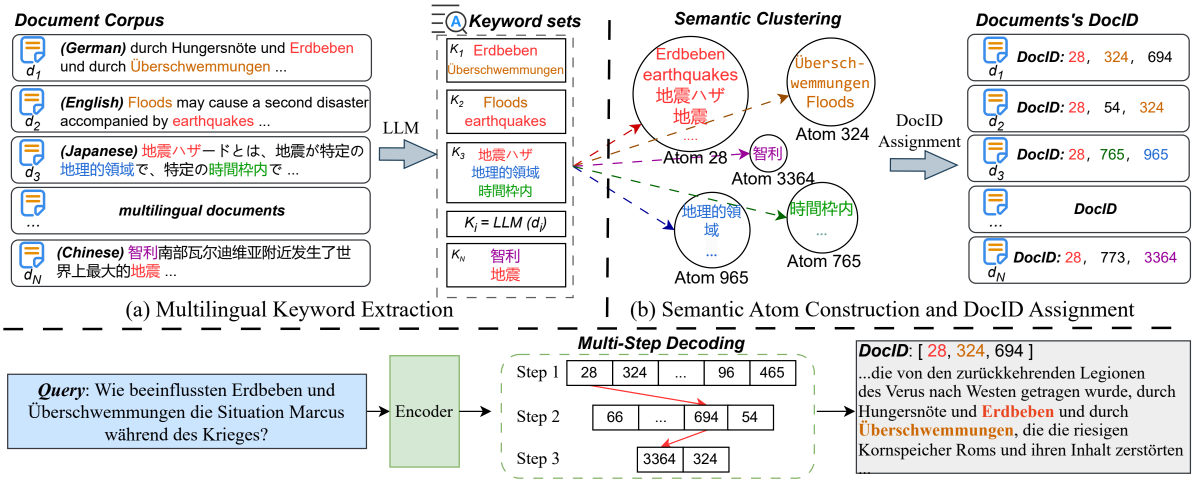
Based on the extracted multilingual keywords, we construct language-independent semantic atoms and assign unique DocIDs to documents.

First, we aggregate all keywords from the entire document collection into a single global set K. Although a specific keyword may appear in the keyword sets K_i of multiple documents, it is represented only once as a distinct element in the global set K. Formally, the global keyword set K is defined as the union of the individual document keyword sets K_i for all $d_i \in \mathcal{D} = [d_1, \dots, d_N]$:

$$K = \bigcup_{i=1}^{N} K_i, \qquad (2)$$

where set K contains n unique keywords, and $n = |K|$ is the total number of distinct keywords in the collection. Let $\{\hat{k}_1, \hat{k}_2, \dots, \hat{k}_n\}$ denote the set of unique keywords. Each keyword $\hat{k}_i$ is encoded into a dense vector representation $v_i \in \mathbb{R}^d$ through a pre-trained text encoder. We construct a similarity matrix $S \in \mathbb{R}^{n \times n}$ where each entry $S_{ij} = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}$ represents the cosine similarity between keywords $\hat{k}_i$ and $\hat{k}_j$.

Keywords are clustered together when their pairwise similarity meets or exceeds a predefined threshold $\theta \in [0, 1]$. Specifically, a fixed number of cluster centers is chosen. Keywords with similarity above the threshold θ are directly assigned to the nearest cluster center, while the others form single clusters. As a result, the process gives C clusters with $C < N$, since similar keywords are grouped together. As the keywords within each cluster are semantically similar, each cluster is


Document Corpus
d1 (German) durch Hungersnöte und Erdbeben und durch Überschwemmungen ...
d2 (English) Floods may cause a second disaster accompanied by earthquakes ...
d3 (Japanese) 地震ハザードとは、地震が特定の地理的領域で、特定の時間枠内で ...
... multilingual documents
dN (Chinese) 智利南部瓦尔迪维亚附近发生了世界上最大的地震 ...
LLM
Keyword sets
K1 Erdbeben Überschwemmungen
K2 Floods earthquakes
K3 地震ハザ 地理的領域 時間枠内
Ki = LLM (di)
KN 智利 地震
Semantic Clustering
Erdbeben earthquakes 地震ハザ 地震 ....
Atom 28
Übersch-wemmungen Floods
Atom 324
智利
Atom 3364
地理的領域 ...
Atom 965
時間枠内 ...
Atom 765
DocID Assignment
Documents's DocID
d1 DocID: 28, 324, 694
d2 DocID: 28, 54, 324
d3 DocID: 28, 765, 965
... DocID
dN DocID: 28, 773, 3364
Multi-Step Decoding
Query: Wie beeinflussten Erdbeben und Überschwemmungen die Situation Marcus während des Krieges?
Encoder
Step 1 28 324 ... 96 465
Step 2 66 ... 694 54
Step 3 3364 324
DocID: [28, 324, 694]
...die von den zurückkehrenden Legionen des Verus nach Westen getragen wurde, durch Hungersnöte und Erdbeben und durch Überschwemmungen, die die riesigen Kornspeicher Roms und ihren Inhalt zerstörten ...

(a) Multilingual Keyword Extraction

(b) Semantic Atom Construction and DocID Assignment

(c) Reasoning Strategy of MGR-CSC

Figure 2: An overview of MGR-CSC's DocID allocation and reasoning. (a) shows how to extract keywords of documents; (b) shows clustering is performed based on cross-lingual semantic similarity of keywords, with each cluster represented by an atom and each document assigned a unique DocID; (c) shows MGR-CSC's reasoning strategy, which returns the identifier corresponding to the query, and narrows the decoding range at each step.

represented by an atom, thereby obtaining a set of atoms $\mathcal{A} = \{a_1, a_2, \ldots, a_c\}$, where a_c is the atomic representation of the c-th cluster.

Finally, since each keyword $\hat{k}_i$ belongs to a cluster represented by an atom a, it can consequently be represented by a. Based on section 3.1, we obtain the set of keywords K_i for each document d_i. Each keyword k_i^m is converted to its atom representation. Let $a_{k_i^m}$ denote the atom representing the keyword k_i^m. The resulting sequence of atom representations is used as the unique DocID for each document d_i,

$$\text{DocID}_i = [a_{k_i^1}, a_{k_i^2}, \ldots, a_{k_i^m}] = [a_1, a_2, \ldots, a_m], \quad (3)$$

where DocID_i represents the DocID of the i-th document. This ensures that all multilingual documents are assigned DocID representations with consistent length and shared semantic space.

3.3 Dynamic Constrained Multi-Step Decoding

In prior methods (Zhuang et al., 2022; Tang et al., 2023), after obtaining document representations with unique DocIDs, the decoding process requires selecting from the complete set of N documents at each step. For a decoding sequence of length m, the method produces a search space scaling as $O(N^m)$. Our proposed method, transforms the retrieval process into a multi-step decoding process of length m within a constrained space comprising c semantic atoms, thereby effectively compressing the decoding space to $O(C^m)$.

To handle the output space, we employ a dynamic constrained multi-step decoding mechanism. During the retrieval process, based on the query q, the model generates the DocID of the target document d_q through the following approach,

$$P(\text{DocID} \mid q) = \prod_{t=1}^{m} P(a_t \mid a_{<t}, q), \quad (4)$$

where P represents the generation probability of the DocID.

The retrieval process generates the target document's DocID through a multi-step decoding procedure under dynamic constraints. At each decoding step t $(1 \leq t \leq m)$, the model predicts the t-th atom a_t based on the query q and the previously generated prefix $\text{DocID}_{<t} = [a_1, \ldots, a_{t-1}]$. The candidate atom set $\mathcal{A}_t$ is defined as:

$$\mathcal{A}_t = \left\{ a_{k_i^t} \,\middle|\, k_i^t \in \text{Constraint}(K_i) \right\}, \quad (5)$$

where $\text{Constraint}(K_i)$ represents the range of documents available under the prefix constraint, and the prefix $\text{DocID}_{<t} = [a_1, a_2, \ldots, a_{t-1}]$. The optimal atom at step t is selected via:

$$a_t = \underset{a_{k_i^t} \in \mathcal{A}_t}{\arg\max}\, P(a_{k_i^t} \mid a_1, \ldots, a_{t-1}, q), \quad (6)$$

with $\text{DocID}[t] = a_t$. Through m iterations, the complete DocID is obtained as:

$$\text{DocID} = [a_1, a_2, \ldots, a_m]. \quad (7)$$

Algorithm 1 Dynamic Constrained Multi-Step Decoding

```
Input Query $q$, Keyword set $K$, Documents $\mathcal{D} = \{d_1, \ldots, d_N\}$, Atom set $\mathcal{A} = [a_1, \ldots, a_c]$
Output Target document DocID
Initialize empty DocID sequence: DocID $\leftarrow$ [ ]
for $t = 1$ to $m$ do
    if $t = 1$ then
        prefix $\leftarrow$ DocID[ ]
    else
        prefix $\leftarrow$ DocID$[1 : t - 1]$
    end if
    $\mathcal{A}_t \leftarrow \{a_{k_i^t} \mid k_i^t \in \text{Constraint}(K_i)\}$
    Compute decoding distribution:
        $P_t \leftarrow \{P(a_{k_i^t} \mid \text{prefix}, q) \mid a_{k_i^t} \in \mathcal{A}_t\}$
    Select optimal atom:
        $a_t \leftarrow \arg\max(P_t)$
    Update DocID: DocID$[t] \leftarrow a_t$
end for
Return DocID $= [a_1, \ldots, a_m]$
```

Algorithm 1 outlines the decoding process applying this dynamic constraint.

4 Experiment

4.1 Datasets

To comprehensively evaluate the model's performance in multilingual document retrieval, we conduct comparative analyses leveraging both standard public benchmarks and our newly constructed multilingual dataset. Below we provide overview of these evaluation datasets:

mMarco100K [1] is a multilingual retrieval benchmark constructed through neural machine translation of the original English MS MARCO dataset (Bonifacio et al., 2021), covering more than 30 languages, consists of a document and a question-answer pair. We randomly sampled non-parallel corpus data in 7 languages, with about 15k data in each language, and a total of about 100k data. Among them, the data set is divided into 6.5k data as a validation set and the rest as a training set.

mNQ320K is a novel multilingual retrieval dataset developed in this study to overcome the limitations of existing resources in low and medium resource language scenarios. It consists of query and document pairs, including about 307K training data and 8K verification data, constructed through a systematic translation methodology following the mMARCO framework. Specifically, We create an extended version of the NQ320K (Kwiatkowski et al., 2019) dataset by translating the original data into seven medium-resource languages spanning diverse language families: Afrikaans (af), French (fr), Arabic (ar), Hindi (hi), Macedonian (mk), Swedish (sv), and Vietnamese (vi).

4.2 Baselines

We conduct comparisons with both traditional retrieval approaches and recent multilingual generative retrieval models. To ensure fairness, we reproduce known advanced generative retrieval methods capable of handling multilingual retrieval.

BM25 (Robertson et al., 2009) represents a standard sparse retrieval model that leverages inverted index structures and operates based on exact keyword correspondence.

LaBSE (Feng et al., 2022) a multilingual sentence encoder that supports 109 languages and maps text from different languages into a unified vector space for cross-lingual retrieval tasks.

mColBERT (Khattab and Zaharia, 2020) a multilingual version of ColBERT that employs late interaction mechanisms for dense retrieval.

ColBERT-xm (Louis et al., 2024) a cross-lingual interaction-based retrieval model that enhances multilingual retrieval through fine-grained token-level interactions.

DSI (Tay et al., 2022) a generative retrieval method that treats documents as training input and constructs DocIDs using hierarchical clustering-based approach.

DSI-QG (Zhuang et al., 2022) a generative retrieval method that trains a query generation model, using short queries to represent the original documents and random numbers to represent DocIDs.

SE-DSI (Tang et al., 2023) a generative retrieval method that utilizes strings containing semantic information as DocIDs. In this study, the titles of multilingual documents are used as the DocIDs for this method.

4.3 Detailed Implementation

In our experiments, all methods were reproduced on the mT5-base model [2] based on the Transformer architecture. Following the work of previous researchers (Zhuang et al., 2022), our training data adheres to the approach of representing documents with multilingual queries. Specifically, the model used for all pseudo-query generation tasks is Llama3.1-8B [3] (Grattafiori et al., 2024)

[1] https://github.com/unicamp-dl/mMARCO

[2] https://huggingface.co/google/mt5-base

[3] https://huggingface.co/meta-llama/Llama-3.1-8B

	Method	en ⇒ oth		fr ⇒ oth		de ⇒ oth		it ⇒ oth		es ⇒ oth		ja ⇒ oth		zh ⇒ oth		AVG	
		@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10
mMarco 100k	BM25	17.53	35.34	11.45	26.48	6.50	15.93	10.48	22.85	14.01	28.87	0.00	0.00	0.58	1.17	8.65	18.66
	Colbert-xm	62.11	83.45	45.45	70.66	36.62	58.76	38.09	61.12	36.00	64.11	41.19	69.26	39.41	65.63	42.69	67.57
	mColbert	52.42	73.17	40.71	68.17	39.98	63.45	38.55	61.84	35.05	63.23	36.49	62.38	33.33	58.55	39.50	64.40
	LaBSE	58.04	79.97	51.10	76.71	42.29	66.71	46.45	70.67	48.03	74.92	39.22	65.02	40.73	70.22	46.55	72.03
	DSI	10.96	23.52	10.01	23.83	9.16	21.34	9.20	21.84	9.58	22.01	9.43	20.50	8.75	20.54	9.50	21.80
	DSI-QG	74.72	87.67	66.05	83.46	61.58	79.96	65.02	82.09	66.90	83.94	62.09	80.76	60.77	78.46	65.21	82.34
	SE-DSI	**76.57**	87.05	**71.48**	83.58	67.15	79.40	68.06	82.18	70.62	84.23	66.35	79.91	67.03	80.87	69.49	82.43
	Ours	72.07	**91.27**	68.53	**87.74**	**70.73**	**90.79**	**69.37**	**89.01**	**71.74**	**91.60**	**68.68**	**88.32**	**67.37**	**86.13**	**69.78**	**89.26**
	Method	af ⇒ oth		fr ⇒ oth		ar ⇒ oth		hi ⇒ oth		mk ⇒ oth		sv ⇒ oth		vi ⇒ oth		AVG	
		@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10
mNQ 320k	BM25	11.73	22.01	12.41	25.45	11.49	24.98	10.75	23.13	11.39	22.48	13.41	26.94	11.84	25.41	11.77	24.16
	Colbert-xm	13.01	28.19	17.26	35.47	12.48	22.75	14.64	29.99	13.73	29.20	20.49	37.00	17.36	35.55	15.57	31.16
	mColbert	17.79	36.78	18.86	38.67	14.30	27.80	14.05	27.49	15.27	30.35	22.68	43.01	18.82	36.52	17.39	34.37
	LaBSE	22.14	45.13	21.96	46.25	14.40	31.40	20.47	42.48	21.44	42.78	25.18	51.65	**24.13**	46.32	21.29	43.29
	DSI	0.11	0.46	0.10	0.41	0.00	0.00	0.00	0.20	0.10	0.40	0.00	0.34	0.10	0.60	0.05	0.33
	DSI-QG	24.34	45.03	**25.08**	48.50	**20.95**	39.48	16.15	37.34	19.84	42.17	25.90	52.03	20.76	42.22	21.32	43.05
	SE-DSI	14.17	26.51	20.62	36.06	15.42	27.06	15.84	28.05	15.35	28.32	22.52	36.49	17.27	33.33	16.76	29.96
	Ours	**26.74**	**49.71**	24.97	**50.47**	20.31	**44.39**	**21.04**	**42.57**	**23.58**	**48.76**	**28.80**	**52.88**	23.38	**47.66**	**24.11**	**48.06**

Table 1: Performance at Recall@1 and Recall@10 on the mMARCO100K and mNQ320K under cross-lingual retrieval settings. Bolded values indicate the best performance among all comparison methods.

with a temperature of 0.7, each document sample is converted into 10 multilingual pseudo-queries through model generation. For keyword generation, the model employed is Llama3.1-8B with a temperature of 0. The model used for semantic similarity calculation is paraphrase-multilingual-MiniLM-L12-v2 [4] (Reimers and Gurevych, 2019).

Training The training was implemented with PyTorch (Paszke, 2019) and Transformers (Vaswani et al., 2017). For mMarco100k, we used a learning rate of 2×10^{-4}, batch size 128, 50 epochs, and m=3 keywords; for mNQ320k, the learning rate was 5×10^{-4}, with the same batch size and m, trained for 100 epochs. The cross-entropy loss function is employed as the objective function. All experiments ran on eight NVIDIA A40 GPUs with 46GB.

Evaluation Metric Aligning with previous studies, we evaluate our model's performance on the validation sets of both datasets, employing Recall@1 and Recall@10 as evaluation metrics. These metrics indicate the fraction of relevant documents retrieved within the top 1 and top 10 positions. Due to the multilingual composition of the candidate document corpus, we present retrieval results for each query language individually.

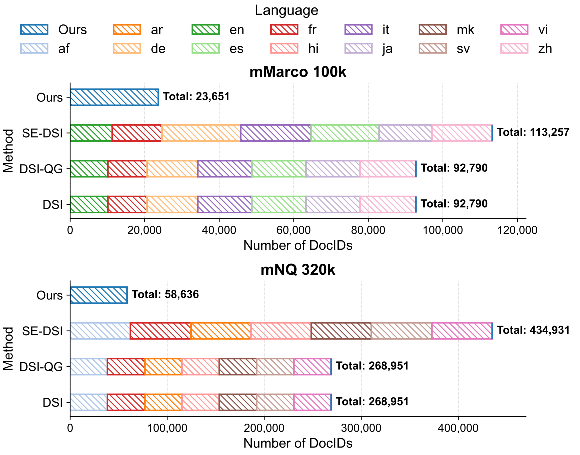


Figure 3: Distribution of the number of DocIDs across different languages for various methods on the mMarco100k and mNQ320k datasets. For SE-DSI, DSI-QG, and DSI, the stacked bar segments represent the distribution of retrieved DocIDs across languages. The overall DocID count for each method is indicated to the right of the corresponding bar.

4.4 Result on mMarco100k and mNQ320k

To demonstrate the performance of our model, we conduct a comparative analysis with existing methods. Table 1 shows retrieval results on two multilingual benchmarks. The mMarco100k dataset, sparse methods such as BM25 perform poorly across languages tasks, especially on non-Latin scripts such

[4] https://huggingface.co/sentence-transformers

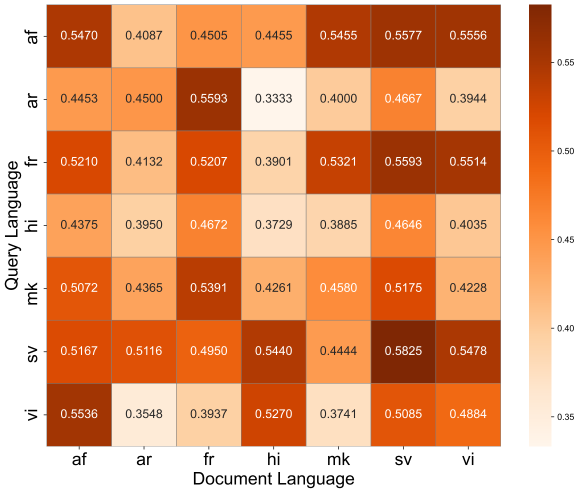


Figure 4: Recall@10 performance of target-language document retrieval with varying source query languages in the mNQ320k dataset

as Japanese and Chinese. This underscores the constraints of lexical matching in languages with morphological and script diversity.

Dense retrieval methods such as LaBSE are more stable but less effective at capturing cross-lingual semantics than generative approaches. DSI and DSI-QG improve Recall@10 on English but underperform on syntactically varied languages, reflecting limited generalization.

MGR-CSC demonstrates consistent performance across all tested languages on mMarco100k dataset. While its Recall@1 on English and French is slightly below SE-DSI, it achieves the highest Recall@10 in nearly all languages, including German, Italian, Spanish, Japanese, and Chinese. This suggests better semantic coverage and decoding stability in multilingual contexts.

The mNQ320k dataset, which includes a broader range of languages, many with fewer resources, offers a more challenging setting. BM25 and DSI exhibit limited retrieval ability in these scenarios. LaBSE maintains moderate performance but shows sensitivity to linguistic variability. DSI-QG and SE-DSI improve cross-lingual recall balance over prior methods but exhibit resource-dependent performance. SE-DSI shows accelerated degradation under low-resource conditions, particularly in linguistically diverse environments.

In contrast, MGR-CSC yields stable and high recall on all languages in the mNQ320k dataset, including significant improvements in low and mid resource languages such as Arabic, Hindi, and Macedonian. These findings are further supported by the results presented in figure 4, which shows the Recall@10 results for the all target language documents across different query languages. The experimental results indicate that MGR-CSC successfully maintains retrieval consistency when processing linguistically diverse data and cross-domain scenarios. Furthermore, the framework demonstrates strong generalizability across both typologically similar and distinct language families.

4.5 Quantitative Analysis of DocID Usage

To quantitatively compare the decoding range between our proposed methodology and existing GIR approaches, we conducted a comprehensive analysis. As depicted in Figure 3, under identical dataset conditions, our method consistently achieves the most restricted decoding space. This significant reduction in the output range substantially enhances decoding efficiency and scalability, rendering our approach particularly advantageous for large-scale document retrieval.

Furthermore, this reduced decoding space is crucial for addressing challenges in multilingual GIR settings. For instance, a single word in high-resource languages, such as English, French, which is typically encoded using one or two tokens. In contrast, in low-resource languages, words are commonly represented as a sequence of subword units. This phenomenon leads to an expanded decoding space and an increased number of decoding steps in low-resource scenarios. To effectively mitigate this, our methodology introduces atomic integer IDs designed to represent clusters of semantically similar keywords. By compressing lexical-level variations into a semantically aligned ID space, this approach effectively minimizes the decoding range and ensures more consistent decoding behavior across linguistic boundaries.

4.6 Ablation Study

To illustrate the contribution of different components, we conducted ablation studies on semantic compression and the decoding strategy separately using the mNQ320k dataset.

Method	R@1	R@10
MGR-CSC	**24.11**	**48.06**
w/o decoding strategy	13.50	31.80
w/o semantic compression	15.58	38.76

Table 2: Performance at Recall@1 and Recall@10 on the mNQ320K under different ablation settings.

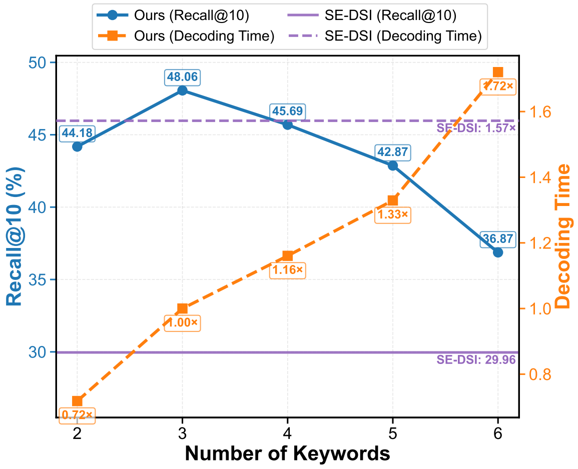


Figure 5: Performance in Recall@10 and decoding time under varying keyword quantities m.

Experimental results (measured by AVG scores) show that removing either component leads to a significant decline in performance. Specifically, removing the decoding strategy leads to a 10.61% drop in Recall@1 and a 16.26% drop in Recall@10. Similarly, removing semantic compression results in 8.53% drop in Recall@1 and 9.30% drop in Recall@10. These results confirm that both semantic compression and the decoding strategy play crucial roles in enhancing retrieval effectiveness.

4.7 The performance of the number of keywords

To examine how the keyword number m affects retrieval performance and decoding time, we experimented on the mNQ320k dataset. As depicted in Figure 5, increasing the number of keywords enhances semantic representation. However, this also increases the length of DocIDs, which consequently slows down decoding and impairs overall performance. The best Recall@10 obtained was 48.06% with three keywords. Fewer keywords, such as two, restrict semantic coverage. In contrast, more keywords, for instance five or six, decrease performance due to increased decoding complexity.

Decoding time scales with DocID length. As the number of keywords increases from two to six, the decoding time approximately doubles. For comparison, SE-DSI achieves a Recall@10 of 29.96%. Its latency is comparable to that of our method when employing the longest DocIDs using six keywords. This similarity emphasizes the strong correlation between sequence length and inference time.

4.8 The performance of the semantic similarity threshold

To assess the sensitivity of keyword clustering to semantic similarity thresholds θ, we evaluated retrieval performance on the mNQ320k dataset with thresholds from 0.5 to 0.9. As illustrated in Figure 6, increasing the threshold from 0.5 to 0.8 resulted in a steady improvement in both Recall@1 and Recall@10, indicating that finer-grained clusters enhance retrieval precision.

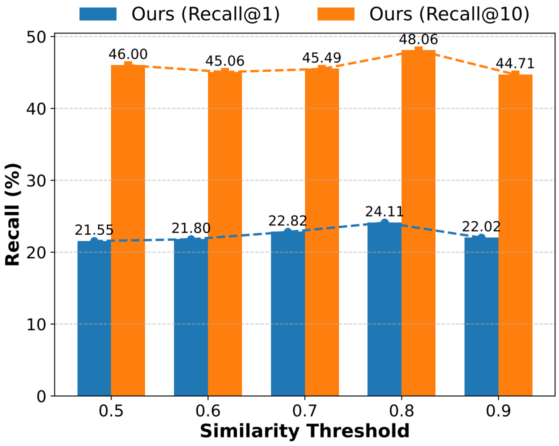


Figure 6: Recall@1 and Recall@10 performance under varying semantic similarity thresholds θ.

The best results were achieved at 0.8, with Recall@1 of 24.48% and Recall@10 of 47.89%. However, raising the threshold to 0.9 caused a performance drop, with Recall@1 decreasing by 2.46% to 22.02% and Recall@10 decreasing by 3.18% to 44.71%. This suggests that an excessively high threshold leads to overly fine-grained clustering, splitting semantically related keywords into separate groups. This reduces semantic generalization and limits the model's retrieval coverage.

4.9 Case Study

To demonstrate the effectiveness of our proposed multilingual generative retrieval method, we present a case study on Swedish-Vietnamese cross-lingual retrieval.

Given the Swedish query *"Vad är Australiens huvudstad?"* (What is the capital of Australia?), the model performs multi-step decoding based on clustered semantic atoms. Each decoding stage progressively narrows the candidate space by focusing on specific semantic dimensions, first detecting the country entity, then identifying the question type, and finally locating the target concept.

Type	Document
Theme	(vi) Danh sách các thủ đô của Úc (List of Australian capital cities)
Content	(vi) Úc có tám thành phố, mỗi thành phố đóng vai trò là trụ sở chính quyền của một tiểu bang hoặc vùng lãnh thổ. Úc được thành lập vào năm 1901. Năm 1927 , trụ sở chính quyền quốc gia đã được di dời và chuyển đến thành phố mới, nơi vẫn tiếp tục đóng vai trò là thủ đô quốc gia cho đến ngày nay. Mỗi thủ đô đều có chức năng tư pháp, hành chính và hành chính. ...
Keywords	Úc , thủ đô , Năm 1927 (Australia, capital, 1927)
Atom set	46, 1788, 14920
Query	(sv) Vad är Australiens huvudstad?
DSI-QG	DocID: 92980 (×)
SE-DSI	DocID: (fr) Territoire de la capitale australienne (×)
Ours-step1	DocID: 46
Ours-step2	DocID: 46, 1788
Ours-step3	DocID: 46, 1788, 14920
Output	46, 1788, 14920

Table 3: Case study on mNQ320k.A Vietnamese document is represented by clustered keyword atom sets. For a given Swedish query, the DocID undergoes a step-wise semantic decoding process along semantic dimensions to retrieve the target document.

The final DocID is composed of shared semantic atoms, enabling successful retrieval of the corresponding Vietnamese document: *"Danh sách các thủ đô của Úc "* (List of Australian capital cities).

5 Conclusion

This paper introduces MGR-CSC, a multilingual generative retrieval method leveraging cross-lingual semantic compression. This method employs semantic clustering to reduce multilingual DocIDs and narrow the decoding space, and applies multi-step constrained decoding to restrict DocID generation. The experimental reveals that our method consistently exhibit outstanding performance compared to existing retrieval approaches when applied to multilingual datasets.

Limitations

Since our model is based on multilingual PLM, its multilingual document understanding capability is consequently limited by the capabilities of this base model. This limitation is particularly pronounced in the context of low-resource languages.

Furthermore, PLMs are mainly designed for text processing. Our existing framework has limited capacity for multimodal information requiring integration of diverse data modalities. It can process text components, yet lacks the inherent ability to understand or reason about cross-modality relationships, thereby restricting its performance in complex multimodal scenarios.

Ethics Statement

This paper proposes a multilingual generative information retrieval method and conducts experiments on both public datasets and extended datasets. Therefore, there are no data privacy implications in this scenario.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant Nos. U21B2027, 62266027, U23A2038), the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202402AG050007, 202302AD080003, 202303AP140008, 202502AD080016), the General Projects of Basic Research in Yunnan Province (Grant Nos. 202301AT070471, 202301AS070047).

References

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–5. IEEE.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Ganesh Chandra and Sanjay K Dwivedi. 2017. Assessing query translation quality using back translation in hindi-english clir. *International Journal of Intelligent Systems and Applications*, 9(3):51.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Sanjay K Dwivedi and Ganesh Chandra. 2016. A survey on cross-language information retrieval. *International Journal on Cybernetics & Informatics (IJCI) Vol*, 5.

Bilel Elayeb, Wiem Ben Romdhane, and Narjes Bellamine Ben Saoud. 2018. Towards a new possibilistic query translation tool for cross-language information retrieval. *Multimedia Tools and Applications*, 77:2423–2465.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. 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*, pages 39–48.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Juntao Li, Chang Liu, Jian Wang, Lidong Bing, Hongsong Li, Xiaozhong Liu, Dongyan Zhao, and Rui Yan. 2020. Cross-lingual low-resource set-to-description retrieval for global e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8212–8219.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. *arXiv preprint arXiv:2402.15059*.

Douglas W Oard, Daqing He, and Jianqiang Wang. 2008. User-assisted query translation for interactive cross-language information retrieval. *Information Processing & Management*, 44(1):181–211.

A Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, and Jun Zhao. 2024a. Multilingual knowledge graph completion from pretrained language models with knowledge constraints. *arXiv preprint arXiv:2406.18085*.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024b. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems*, 36:46345–46361.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2023. Semantic-enhanced differentiable search index inspired by learning strategies. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4904–4913.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, and 1 others. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023. Novo: Learnable and interpretable document identifiers for model-based ir. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2656–2665.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, and 1 others. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. 2019. Robust document representations for cross-lingual information retrieval in low-resource settings. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 12–20.

Puxuan Yu and James Allan. 2020. A study of neural matching models for cross-lingual ir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1640.

Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. 2022. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4345–4353.

Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Fangchao Liu, and Zhao Cao. 2024. Generative retrieval via term set generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 458–468.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, Neha Verma, William Hu, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. *arXiv preprint arXiv:1906.03492*.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128*.

A Appendix

A.1 The results of Recall@100 on the mNQ320k dataset

Method	af ⇒ oth	fr ⇒ oth	ar⇒ oth	hi⇒ oth	mk⇒ oth	sv⇒ oth	vi⇒ oth	AVG
BM25	34.22	38.64	37.19	35.99	35.48	40.34	38.55	37.20
Colbert-xm	43.20	52.80	35.40	48.10	48.48	54.77	50.25	47.57
mColbert	52.58	56.15	44.56	42.09	46.44	58.48	52.60	50.41
LaBSE	65.51	**67.48**	51.35	**64.13**	65.51	**70.47**	**65.88**	64.33
DSI-QG	63.04	66.39	57.84	60.56	64.43	67.07	61.98	63.04
SE-DSI	38.45	49.70	39.84	40.35	43.84	50.61	44.43	43.88
Ours	**67.49**	67.11	**63.12**	62.79	65.72	69.23	64.14	**65.66**

Table 4: Performance at Recall@100 on the mNQ320K. Bolded values indicate the best performance among all comparison methods.

To demonstrate the performance of our method on high-ranks@100, we extend the recall@100 results on the mNQ320k dataset. Table 4 presents the results, showing that our method outperforms others in most languages as well as in the average score, with the AVG exceeding the current baseline by 1.33%.