# Steering Large Language Models for Cross-lingual Information Retrieval

Ping Guo[*]
Institute of Information Engineering,
Chinese Academy of Sciences &
School of Cyber Security, University
of Chinese Academy of Sciences
Haidian, Beijing, China
guoping@iie.ac.cn

Yubing Ren[*]
Institute of Information Engineering,
Chinese Academy of Sciences &
School of Cyber Security, University
of Chinese Academy of Sciences
Haidian, Beijing, China
renyubing@iie.ac.cn

Yue Hu[†]
Institute of Information Engineering,
Chinese Academy of Sciences &
School of Cyber Security, University
of Chinese Academy of Sciences
Haidian, Beijing, China
huyue@iie.ac.cn

Yanan Cao
Institute of Information Engineering,
Chinese Academy of Sciences &
School of Cyber Security, University
of Chinese Academy of Sciences
Haidian, Beijing, China

Yunpeng Li
Institute of Information Engineering,
Chinese Academy of Sciences &
School of Cyber Security, University
of Chinese Academy of Sciences
Haidian, Beijing, China

Heyan Huang
Beijing Institute of Technology
Haidian, Beijing, China

## ABSTRACT

In today's digital age, accessing information across language barriers poses a significant challenge, with conventional search systems often struggling to interpret and retrieve multilingual content accurately. Addressing this issue, our study introduces a novel integration of applying Large Language Models (LLMs) as Cross-lingual Readers in information retrieval systems, specifically targeting the complexities of cross-lingual information retrieval (CLIR). We present an innovative approach: Activation Steered Multilingual Retrieval (ASMR) that employs "steering activations"—a method to adjust and direct the LLM's focus—enhancing its ability to understand user queries and generate accurate, language-coherent responses. ASMR adeptly combines a Multilingual Dense Passage Retrieval (mDPR) system with an LLM, overcoming the limitations of traditional search engines in handling diverse linguistic inputs. This approach is particularly effective in managing the nuances and intricacies inherent in various languages. Rigorous testing on established benchmarks such as XOR-TyDi QA, and MKQA demonstrates that ASMR not only meets but surpasses existing standards in CLIR, achieving state-of-the-art performance. The results of our research hold significant implications for understanding the inherent features of how LLMs understand and generate natural languages, offering an attempt towards more inclusive, effective, and linguistically diverse information access on a global scale.

[*] Equal Contribution.
[†] Corresponding Author.

*SIGIR '24, July 14–18, 2024, Washington, DC, USA*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657819

## CCS CONCEPTS

• **Information systems → Multilingual and cross-lingual retrieval**; • **Computing methodologies → Natural language generation**.

## KEYWORDS

Cross-lingual Information Retrieval, Activation Steering, Large Language Models

**ACM Reference Format:**
Ping Guo, Yubing Ren[*], Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering Large Language Models for Cross-lingual Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3626772.3657819

## 1 INTRODUCTION

In the contemporary digital era, information retrieval has become an integral part of daily life, profoundly influencing how we seek and consume information. From online academic research to daily news updates, and from e-commerce browsing to social media interactions, the ability to efficiently access relevant information is pivotal. However, the effectiveness of conventional information retrieval systems is largely constrained by linguistic boundaries. Users often encounter limitations in accessing information due to language discrepancies; a search query in one language can yield vastly different results when translated into another, creating a digital divide in information access. This linguistic boundary not only hampers the global exchange of knowledge but also accentuates the urgent need for advancements in Cross-Lingual Information Retrieval (CLIR). By transcending the linguistic boundary, CLIR promises more inclusive and comprehensive access to information, catering to a diverse, multilingual user base.

Despite its critical importance, CLIR presents a unique set of challenges, primarily stemming from the need to understand and interpret queries and documents across a multitude of languages.

This complexity is compounded by the inherent nuances in linguistic structures, idiomatic expressions, and cultural contexts that vary significantly from one language to another. Furthermore, an essential component of this process involves a post-hoc translation, wherein the retrieved documents in various languages must be translated back into the query's language. This step is fraught with potential for propagational errors, as nuances and meanings can be lost or altered in translation, leading to inaccuracies in the retrieved information. Such challenges not only exacerbate the difficulty of developing effective cross-lingual retrieval systems but also highlight the intricacies of language processing that must be addressed to ensure accurate and relevant information retrieval across linguistic barriers.

The advent of Large Language Models (LLMs) has been a game-changer in the field of Natural Language Processing (NLP), offering innovative solutions to a range of complex challenges, including those in CLIR. The application of LLMs in NLP extends beyond basic text processing, encompassing advanced tasks such as semantic analysis, context understanding, and even creative content generation. Specifically, in the context of CLIR, the advantages of LLMs are multifold:

**(1) Cross-Lingual Transferability:** LLMs are trained on extensive, diverse multilingual datasets, enabling them to effectively align and interpret multiple languages. **(2) Natural Language Understanding:** Their advanced architectures allow for a high degree of accuracy in interpreting user queries and extracting relevant information. **(3) Language Generation Capabilities:** LLMs excel in producing grammatically correct and contextually relevant text across different languages. **(4) Common-Sense Knowledge:** A critical aspect of LLMs is their ability to integrate and apply common-sense knowledge, essential for understanding the nuances and implied meanings in both queries and content. These capabilities position LLMs as pivotal tools in revolutionizing CLIR, enabling not just the processing and understanding of queries across languages, but also ensuring the delivery of accurate and contextually relevant information back to the user in their own language.

In this study, we address the challenges of CLIR by leveraging the capabilities of LLMs. Our innovative approach, Activation Steered Multilingual Retrieval (ASMR), integrates a Multilingual Dense Passage Retrieval (mDPR) model with an LLM. The mDPR model is employed to initially retrieve relevant documents, thereby addressing the LLM's limitations in handling extensive input lengths. Then we treat the LLMs as a cross-lingual reader to process and generate responses based on the retrieved results. The key innovation in ASMR is the introduction of "steering activations", which applies designed vectors to shift specific activation within the LLM towards targeted directions, thus controlling the attributes of the final generated response. In the context of CLIR, our focus is on manipulating two key attributes: the accuracy of the response and the coherence of the response language. Our methodology identifies two distinct sets of attention heads within the LLM through different attribute probing classifiers. One set is specialized at distinguishing between languages, and the other excels in accurate question answering. During the inference phase, we strategically shift the activation of these attention heads in two directions: one to improve language detection and the other to enhance the accuracy of the answers. This process is executed autoregressively, continuing until a complete and coherent answer is generated. The usage of steering activation ensures that the answers align closely with the user's query language and accurately address the content of the query.

Our attribute probing classifiers offer deeper insights into the learning mechanisms of the LLM, particularly in how it answers questions and where its cross-lingual transferability stems from. Our findings reveal that approximately half of the attention heads in the LLM demonstrate a clear ability to distinguish between different languages. However, intriguingly, only about 34% of the attention heads in the middle layers of the LLM perform better than a random guess in accurate question answering. To further validate our approach, we conducted extensive experiments assessing the accuracy and relevance of our LLM's generated responses in relation to user queries. These experiments utilized CLIR benchmarks such as XOR-TyDi QA [3], and MKQA [43]. The results from these benchmarks are highly encouraging, indicating that ASMR not only achieves but surpasses state-of-the-art performance when compared against both previous CLIR baselines and other LLM-based methods.

## 2 RELATED WORK

### 2.1 Cross-lingual Information Retrieval

CLIR has been a focal point of research, with various approaches being explored. A significant strand of work has utilized translation models to bridge the language gap in CLIR tasks [40, 42, 56, 72, 77]. These methods typically involve translating multilingual queries to a pivot language, such as English, or translating retrieved documents back to the user's native language for easier comprehension.

The evolution of CLIR research [13, 76, 78, 80] has been significantly influenced by advancements in cross-lingual representations, with models like XLM-R and m-BERT [28, 69] at the forefront. These improvements are often grounded in refined cross-lingual embeddings [17, 73–75]. Additionally, there is a growing body of work [22, 23, 26, 36, 37, 54] that focuses on distilling insights from monolingual models into multilingual frameworks.

Innovative techniques such as code-switching [16, 38, 66], query generation [4, 54, 82], and sequential sentence relation modeling [39, 41, 79] have been employed to tackle the complexities of CLIR. Variational models like VMSST [2, 70] disentangle semantic information, crucial for understanding and retrieving across languages, while contrastive learning [20, 25, 60, 70, 81] aligns semantically similar sentence embeddings. Cross-lingual soft prompts fine-tune models for better CLIR [23]. CLIR's versatility extends to enhancing areas like fact-checking [21]. Enhanced CLIR converges on two fronts: knowledge distillation [22, 23, 36, 54] enriches multilingual architectures, and contrastive learning advances embedding alignment. These developments illustrate CLIR's dynamic landscape, fostering more efficient and accurate retrieval systems across languages.

### 2.2 Large Language Models for Search

The advent of LLMs such as ChatGPT[1] has catalyzed a paradigm shift in NLP. These models are celebrated for their advanced capabilities in language understanding, generation, and reasoning, which have significantly expanded the horizons of generalization

[1] https://chat.openai.com/

within the field. The intersection of LLMs and Information Retrieval (IR) systems is a burgeoning area of research, marked by rapid developments that promise to reshape IR paradigms. LLMs are being integrated into IR systems at various levels, from enhancing query reformulation with sophisticated query rewriters [14, 44, 45, 47, 55, 58, 68] to improving the retrieval phase with intelligent retrievers [11, 61, 83]. Furthermore, they play a pivotal role in refining search results through advanced rerankers [6, 29, 52] and in providing precise answers via sophisticated readers [27, 65]. This paper zeroes in on harnessing the power of LLMs to confront and mitigate the challenges associated with CLIR, thereby enhancing the model's utility in multilingual search contexts.

### 2.3 Activation Steering

Steering a pre-trained LLM's activations offers a minimally invasive way to influence its behavior during inference. Activation editing, described in recent literature [19, 33], subtly guides model outputs without extensive retraining, beneficial for applications like style transfer using pre-trained or manually curated steering vectors [59]. This contrasts with weight editing methods, which aim for minimal model alteration but may reduce overall robustness [9, 18, 24, 48, 49]. An alternative, Contrast-Consistent Search [10], finds truthful directions within activations using logical consistencies, less resource-intensive than reinforcement learning [7, 15, 50]. Activation perturbation, rooted in plug-and-play methods for controllable text generation [12, 31, 35], has led to mechanistic interpretability techniques like Inference Time Intervention (ITI) [34], locating influential attention heads and steering them toward truthful outputs with minimal samples.

Building on the foundations laid by ITI, ASMR seeks to expand the range of controllable attributes, allowing for multi-faceted manipulation of the model's activations. By doing so, we aim to tap into the latent knowledge embedded within the model, enabling richer and more precise control over its generative capabilities.

## 3 ACTIVATION STEERED MULTILINGUAL RETRIEVAL

In this section, we outline the functionality of ASMR, which employs both a retriever model and an LLM. The role of the retriever model is to supply the LLM with documents relevant to the user's query. For this purpose, we utilize an unmodified mDPR in our experiments. The LLM works as the reader, and is tasked with dual objectives: firstly, to extract the most relevant answer from the documents provided by the retriever, and secondly, to generate the final response in the language of the user's query.

We will first briefly describe some key elements of the transformer architecture to set notation and context. Then, we will delve into the specifics of how we guide the LLM to accomplish these objectives. The next subsection focuses on identifying attention heads within the LLM that significantly impact certain attributes, a process achieved using probing classifiers. Following this, the final subsection explains our design of steering activations and their integration into the attention heads of the LLM. This is a crucial step in directing the LLM's output to align with our goals of accuracy in content and coherence in language.

### 3.1 Notations & Preliminaries

Transformer [67] architecture is a revolutionized model in NLP. To bring clarity to the structure of this architecture, we distill the architecture to its essence, the *transformer layers*, indexed by the variable $l$. Each layer comprises two principal modules: the Multi-Head Attention (MHA) mechanism, which enables the model to focus on different parts of the input sequence simultaneously, and the Multi-Layer Perceptron (MLP) layer, which applies a series of nonlinear transformations. At the heart of Transformer architecture is the attention mechanism, which allows the model to weigh the importance of different parts of the input data differently, thus allowing for the parallel processing of different representation subspaces at different positions.

Tokens are initially mapped into a high-dimensional vector space $x_0 \in \mathbb{R}^d$ at the start of inference, initiating the residual stream. This stream unfolds as a sequence of vectors $x_0, \ldots, x_n$, where each transformer layer processes the vector $x_i$, integrates the computations, and generates the subsequent vector $x_{i+1}$ by adding the processed outcome back to $x_i$. The final vector in the residual stream of the last layer is decoded to predict the next token's distribution.

MHA within each layer is characterized into $H$ different heads by $H$ distinct linear operations, which facilitate the model's ability to attend to information from different representation subspaces. The formulation of the MHA operation is as follows:

$$x_{l+1} = x_l + \sum_{h=1}^{H} W_l^h \textsc{Att}_l^h (Q_l^h x_l), \tag{1}$$

where $Q_l^h \in \mathbb{R}^{d_h \times d}$ transforms stream activations into each head dimension, a $d_h$-dimensional head space, and $W_l^h \in \mathbb{R}^{d \times d_h}$ maps it back. The Att function enables the interactions between input tokens, enabling the model to effectively integrate information across the entire input sequence. we perform our analysis and interventions at post the Att operation and prior to applying $W_l^h$, with activations represented as $x_l^h \in \mathbb{R}^d$.

### 3.2 Locating Attention Heads with Probing Classifiers

We aim to pinpoint the attention heads within the network that most effectively influence specific attributes of the generated output. To achieve this, we utilize a technique known as "probing" [1, 8, 63]. Probing involves training a classifier on the network's activations to identify differences in output based on given input characteristics. Specifically, our interest lies in differentiating outputs based on two attributes: **the accuracy of the response** and **the coherence of the response language**. For this purpose, our probe is mathematically represented as $p_\theta(x_l^h) = \text{sigmoid}(\theta^\top x_l^h)$, where $\theta$ is a vector in the space $\mathbb{R}^{d_h}$. Each attention head has an associated probe, with $x_l^h$ indicating the contribution from the $h$-th attention head at layer $l$ to the model's overall output.

***Identifying Attention Heads for Accuracy in Content.*** We investigated the question-answering capabilities of our LLM using the Natural Questions dataset [32], which was originally curated for end-to-end question answering tasks in English. This dataset

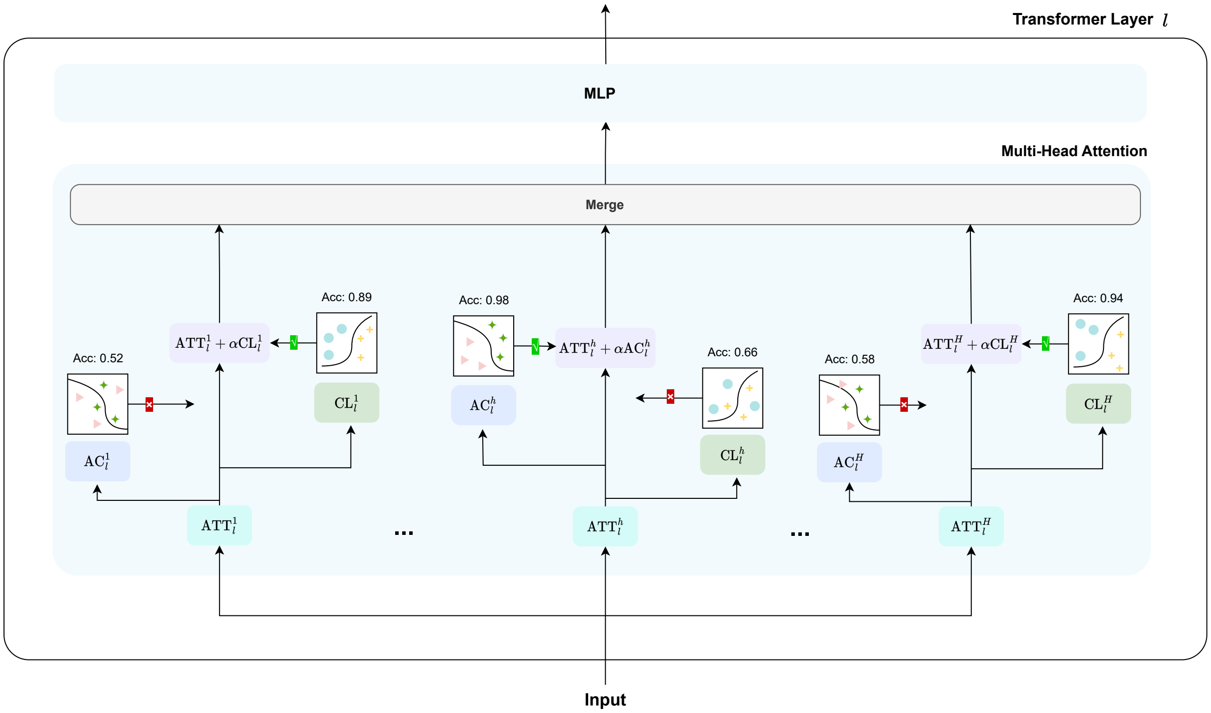


**Figure 1: The overall framework of ASMR. We visualize the $l$-th Transformer layer. Our ASMR first applies two probing classifiers on each attention head for each transformer layer to probe which attention heads affect the most in the two attributes: Coherence in Languages, marked with "CL" in the figure, and Accuracy in Content as "AC" in the figure. We decide the location to apply steering activation based on the accuracy of the probing classifiers. We select top-$K$ heads for each attribute and choose the weight direction of the probing classifier as steering activation.**

comprises real user queries from Google search and their corresponding answers located within Wikipedia articles, annotated by human experts. We selected the validation subset, which includes 7830 entries, each with five-way annotated responses.

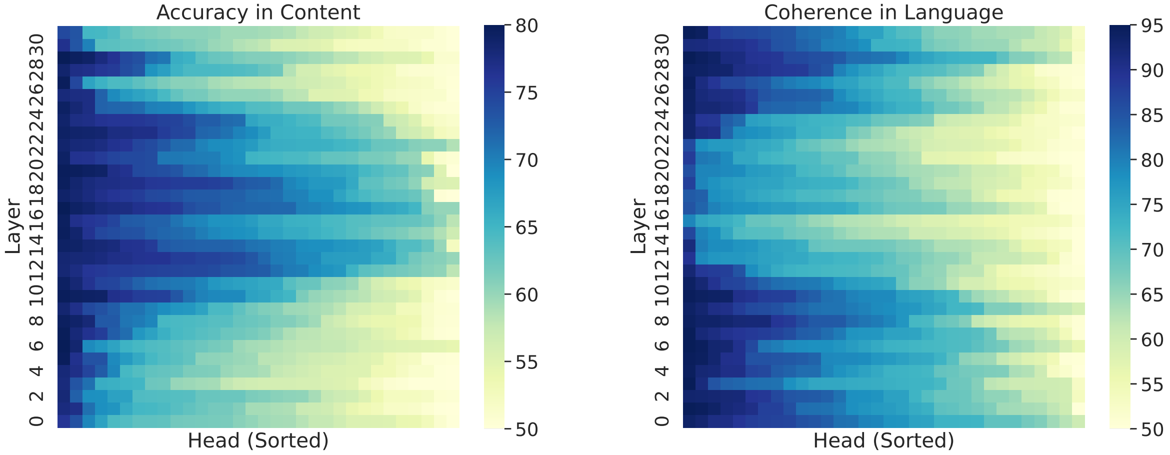
For each question, we created two pairs: one with the question and a document containing the answer, and another with the question and a document without the answer. Each pair received a binary score, $r_i$, indicating its relevance. The LLM processed each pair, and we collected the head activations at the last token to form a probing dataset $\{(x_l^h, r_i)\}_{i=1}^N$ for every attention head across all layers. We divided each resulting dataset into training and validation sets in a 4:1 ratio, applied a binary linear classifier to the training data, and measured each head's contribution to performance using the validation set accuracy.

The results in Figure 2a showcased a notable specialization among the attention heads. While many heads only achieved baseline accuracy, comparable to random chance, others demonstrated substantial predictive power. For instance, the 9th head of the 14th layer stood out with a validation accuracy of 79.3%. A broader analysis, illustrated in Figure 2a, indicated that middle layers predominantly processed the information and a select few heads in each layer exhibited exceptional performance.

***Identifying Attention Heads for Coherence in Language.*** To assess language coherence, we turned to the FLORES-200 benchmark [62], a comprehensive multilingual dataset encompassing translations of 3001 English sentences into 204 languages by professional translators. These sentences, drawn from diverse sources such as WikiNews[2], WikiJunior[3], and WikiVoyage[4], cover a broad range of subjects. From this dataset, we chose ten frequently tested languages, which are German, Spanish, French, Italian, Dutch, Portuguese, Thai, Turkish, Vietnamese, and Chinese, and selected 20 unique sentences for each. Notably, the sentences differed across languages to ensure diversity.

We constructed sentence pairs for probing by concatenating two sentences at random, labeling the pair with '$y = 0$' if both sentences were in the same language, creating 3800 such pairs. To balance the dataset, we also generated 3800 pairs of sentences from different languages, labeled with '$y = 1$'. Similar to our accuracy assessment, we input each sentence pair into the LLM, extracted the activations at the last token, and compiled a probing dataset $\{(x_l^h, y_i)\}_{i=1}^N$ for every head in each layer. Following the same procedure, we split the datasets, trained binary classifiers, and evaluated head-related performance using validation accuracy.

[2] https://en.wikinews.org/wiki/MainPage
[3] https://en.wikibooks.org/wiki/Wikijunior
[4] https://en.wikivoyage.org/wiki/Main_Page



**(a) Accuracy in Content**

**(b) Coherence in Languages**

**Figure 2: Visualization of the probing accuracy for different attributes: accuracy in content and coherence in languages. We show the linear probe accuracy on the validation set for all heads in all layers in Llama2-7B, sorted row-wise by accuracy. Darker blue represents higher accuracy. 50% is the baseline accuracy from random guessing.**

Contrasting with the findings on content accuracy, the coherence in language experiments revealed a different specialization trend. Our results, depicted in Figure 2b, showed that early and top layers were more active in processing language coherence, with middle layers performing close to baseline performance. Remarkably, the 13th head in the 28th layer achieved the highest validation accuracy at 98.5%, indicating its critical role in language coherence tasks.

### 3.3 Designing and Implementing Steering Activations

Our probing experiments have shed light on the mechanisms by which the LLM processes two critical attributes: accuracy in content and coherence in language. These insights have led us to develop a method that can potentially enhance the LLM's performance on benchmark datasets. The cornerstone of this method is strategic intervention during inference—specifically, steering the activations towards the directions that correspond to these attributes, with the anticipation that such adjustments will yield more accurate and coherent responses.

Rather than indiscriminately adjusting all attention heads within the LLM, our approach is more nuanced. We target our interventions to only those heads that, as indicated by the patterns in Figure 2, have a pronounced impact on either accuracy or coherence. By focusing on the top $K$ heads for each attribute, we can finely tune the model's output without disrupting its overall structure.

The methodology for choosing the direction of intervention involves the weight direction gleaned from the linear probing classifier, as discussed in subsection 3.2. Intervening in this manner is akin to taking a gradient descent step specifically tailored to enhance the likelihood of producing coherent or accurate predictions. These calculated deviations guide us in adjusting the model's attention mechanism. The adjusted Multi-Head Attention can be expressed by the following equation:

$$x_{l+1} = x_l + \sum_{h=1}^{H} W_l^h \left( \mathrm{Att}_l^h(Q_l^h x_l) + \alpha \theta_l^h \right), \;\; \theta_l^h \in \{\mathbf{0}, \mathrm{AC}_l^h, \mathrm{CL}_l^h\}, \tag{2}$$

where $\mathrm{AC}_l^h$ represents the weight direction for Accuracy in Content probing classifier at layer $l$ in the $h$-th attention head, while $\mathrm{CL}_l^h$ means the weight direction for Coherence in Language probing classifier. For those heads not selected for intervention, $\theta$ is set to a zero vector, signifying no adjustment. This method is parameterized by two critical values: $K$, the number of heads we choose to intervene upon, and $\alpha$, the magnitude of the intervention. While we have yet to establish a theoretical basis for the optimal values of these parameters, we conduct experimental sweeps to empirically identify their most effective settings. The choice of $\alpha$ is particularly sensitive and should be determined according to the user's priorities—if accuracy is paramount, a higher $\alpha$ should be set. This intervention process is applied autoregressively for subsequent token predictions and is designed to be independent of the decoding algorithm used.

## 4 EXPERIMENTAL SETUP

Our experimental evaluation aims to rigorously test the effectiveness of ASMR in CLIR. We selected two distinct datasets for this purpose: XOR-TyDi QA [3] and MKQA [43]. This dual-dataset approach is critical for comprehensively assessing the versatility and full capability of ASMR in addressing diverse CLIR challenges.

### 4.1 Evaluation Benchmarks

Our evaluation targets two primary tasks: (1) retrieving information from English collections using queries in multiple languages, and (2) generating accurate answers in these languages from English collections. To this end, we utilize two distinct datasets, XOR-TyDi

**Table 1: The performance of previous CLIR methods and LLMs on the MKQA dataset. We report the R@2kt scores and the best performance is marked with bold font. For a fair comparison, we adapt SFT to use only 5% of Natural Questions QA pairs.**

| Models | German | Spanish | French | Italian | Dutch | Portuguese | Thai | Turkish | Vietnamese | Chinese | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Previous CLIR Methods | | | | | |
| BM25+Ext.reader+MT | 43.9 | 45.3 | 41.7 | 41.1 | 45.2 | 46.4 | 45.9 | 42.7 | 44.3 | 38.2 | 43.5 |
| mDPR+Ext.reader+MT | 50.5 | 48.0 | 48.9 | 41.2 | 48.4 | 48.6 | 46.1 | 45.0 | 48.1 | 46.8 | 47.2 |
| CORA | 44.6 | 45.3 | 44.8 | 44.2 | 47.3 | 40.8 | 45.0 | 34.8 | 33.9 | 33.5 | 41.4 |
| Sentri | 56.5 | 55.9 | 55.1 | 54.3 | 56.3 | **54.8** | 55.3 | 53.0 | 54.4 | 50.2 | 54.6 |
| | | | | | | LLM-based Methods | | | | | |
| BLOOM-7B | | | | | | | | | | | |
| +Supervised Fine-tuning | 49.3 | 46.9 | 46.7 | 48.8 | 50.1 | 37.0 | 38.8 | 39.5 | 37.9 | 52.1 | 44.7 |
| +Few-shot Prompting | 52.3 | 48.7 | 47.6 | 51.2 | 50.9 | 40.5 | 39.0 | 41.3 | 37.4 | 53.1 | 46.1 |
| +ASMR | 53.1 | 48.7 | 49.2 | 53.5 | 54.0 | 41.3 | 43.0 | 42.4 | 46.1 | **55.6** | 48.7 |
| LLAMA2-7B | | | | | | | | | | | |
| +Supervised Fine-tuning | 54.5 | 57.2 | 55.7 | 54.0 | 54.6 | 45.3 | 51.1 | 51.8 | 51.2 | 50.7 | 52.6 |
| +Few-shot Prompting | 56.0 | 56.4 | **58.6** | 53.3 | 56.5 | 48.1 | 54.0 | **55.2** | 53.3 | 51.5 | 54.4 |
| +ASMR | **59.0** | **57.8** | 58.0 | **56.2** | **56.8** | 46.4 | **56.0** | 54.6 | **56.3** | 51.5 | **55.4** |

QA and MKQA, which offer diverse challenges in terms of collection size, relevance, and language variety.

**XOR-TyDi QA:** The XOR-TyDi QA dataset [3] is designed for multilingual open-domain question answering. It includes questions in 7 different languages, originally sourced from TYDI QA, and answered using Wikipedia data. The answers are either in the same language as the question or translated from English. We assess our method on this dataset using the F1 score, following standard evaluation practices [57].

**MKQA:** The MKQA dataset [43] is an extensive multilingual benchmark for open-domain question answering. It contains over 10,000 examples with each question translated into 26 languages. For our assessment, we select a subset of 10 languages to evaluate our method. This dataset is particularly challenging due to the diversity of answer types it includes, ranging from numeric data to short phrases. MKQA's broad language coverage makes it an ideal benchmark for testing our method's ability to adapt and deliver accurate answers across different languages.

## 4.2 Evaluation Metrics

Our approach to evaluating generation quality is aligned with established methodologies in the field [57]. For the MKQA dataset, we primarily use the recall scores for the initial 2000 tokens (R@2kt) to assess the performance. This metric effectively captures the relevance of generated responses in the context of large-scale information retrieval.

For the XOR-TyDi QA task, our evaluation encompasses three distinct metrics: F1, Exact Match (EM), and BLEU scores. The F1 metric evaluates the token-level overlap between the generated answer and the gold standard, providing a measure of precision and recall. EM is a stricter criterion, assessing whether the system's output exactly matches the correct answer. The BLEU score, as defined by [51], quantifies the similarity between the model-generated output and the gold standard by comparing their n-grams. To ensure the reliability of our results, we confirm statistical significance through a two-tailed paired $t$-test, with a p-value threshold set at 0.05.

## 4.3 Compared Baselines

In our evaluation, we compare our method against a diverse range of baselines, encompassing previous CLIR methods, different LLMs like BLOOM [71] and Llama2-7B [64], and various LLM adaptation techniques such as supervised fine-tuning and few-shot prompting.

**Previous CLIR Methods.** Our first set of baselines includes traditional methods in CLIR, which generally involve a combination of retrieval models and reader components. We consider the following:

**(1) BM25 + Ext.reader + MT:** This method employs a Neural Machine Translation (NMT) model to translate queries into English, followed by a monolingual neural retrieval (like ColBERT [30]) for English-to-English query-document matching. The pipeline concludes with BM25 [46] as the retrieval model, an extractive reader, and machine translation for final response generation. **(2) mDPR + Ext.reader + MT:** Similar to the above but using mDPR, this approach extracts answers using a multilingual dense passage retrieval system, then applies an extractive reader model, and concludes with answer translation. It assesses the effectiveness of multilingual generation models compared to extractive reading combined with translation. **(3) CORA:** As a unified many-to-many QA model, CORA [5] answers questions across multiple languages, including those without specific language data. It uses a new dense passage retrieval algorithm for CLIR combined with a multilingual autoregressive generation model, avoiding the need for intermediate translation or retrieval steps. **(4) Sentri:** Developed by [57], Sentri utilizes a single encoder for both query and passage retrieval from a multilingual collection, paired with a cross-lingual generative reader. This method represents a strong traditional supervised approach in CLIR before the advent of LLM-based techniques.

**Table 2: Performance of previous CLIR methods and LLMs on XOR-TyDi QA dataset. We report the F1, EM, and BLEU scores and mark the best performance with bold font. For a fair comparison, we adapt SFT to use only 5% of Natural Questions QA pairs.**

| Method | F1 | | | | | | | Macro Average | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Arabic | Bengali | Finnish | Japanese | Korean | Russian | Telugu | F1 | EM | BLEU |
| Previous CLIR Methods | | | | | | | | | | |
| BM25+Ext.reader+MT | 9.2 | 15.8 | 14.4 | 4.8 | 7.9 | 5.2 | 0.5 | 8.3 | 4.6 | 7.5 |
| mDPR+Ext.reader+MT | 18.9 | 11.2 | 21.1 | 3.9 | 10.6 | 8.1 | 13.9 | 12.5 | 7.7 | 13.5 |
| CORA | 42.9 | 26.9 | 41.4 | 36.8 | 30.4 | 33.8 | 30.9 | 34.7 | 25.8 | 23.3 |
| Sentri | 52.5 | 31.2 | 45.5 | 44.9 | 43.1 | 41.2 | 30.7 | **41.3** | 34.9 | 30.7 |
| LLM-based Methods | | | | | | | | | | |
| BLOOM-7B | | | | | | | | | | |
| +Supervised Fine-tuning | 44.5 | 51.0 | 19.2 | 41.2 | 14.5 | 32.6 | 39.1 | 34.6 | 27.5 | 21.3 |
| +Few-shot Prompting | 42.6 | **55.1** | 17.5 | 40.3 | 8.6 | 33.8 | 40.9 | 34.1 | 28.9 | 24.2 |
| +ASMR | 44.2 | 53.9 | 20.1 | 42.6 | 10.1 | 36.1 | **42.5** | 35.6 | 30.4 | 27.4 |
| LLAMA2-7B | | | | | | | | | | |
| +Supervised Fine-tuning | 49.7 | 32.0 | 40.2 | 45.2 | 49.7 | 41.8 | 8.5 | 38.2 | 32.4 | 29.9 |
| +Few-shot Prompting | 50.8 | 31.0 | **50.6** | 42.1 | 49.3 | 42.6 | 7.6 | 39.1 | 33.8 | 30.5 |
| +ASMR | **55.2** | 36.4 | 43.8 | **47.0** | **51.7** | **44.6** | 10.5 | **41.3** | **35.1** | **31.0** |

**Large Language Models.** To evaluate the impact of ASMR on various LLMs, we conducted experiments using two mainstream open-source models with different parameters:

**(1) BLOOM-7B:** This multilingual model [71] covers 46 natural languages and 13 programming languages. It is a decoder-only Transformer model trained on the ROOTS corpus. We selected the BLOOM-7.1B variant for its ability to perform competitively across diverse benchmarks, especially after multitask prompted fine-tuning. **(2) Llama2-7B:** Released by MetaAI, this model [64] comes in several versions with varying parameters. The Llama2-7B model was chosen for its distinct size compared to BLOOM-7B, offering a comprehensive comparison. It is trained on a large corpus, providing robust performance across multiple tasks.

**LLM Adaptation Methods.** Our method, ASMR, which does not necessitate fine-tuning of LLM parameters, is compared against other LLM application methods like supervised fine-tuning and few-shot prompting.

**(1) Supervised Fine-tuning (SFT):** Involves training the model on QA pairs to generate relevant answers and discourage irrelevant responses. This method alternates between QA pair training and pretraining on Open Web Text [53], fine-tuning all model parameters as suggested by prior research. For a fair comparison with few-shot prompting and ASMR, we adapt SFT to use only 5% of Natural Questions QA pairs. **(2) Few-shot Prompting (FSP):** This approach uses a single example (1-shot) to demonstrate the task to the LLM. Given the long input nature of CLIR tasks, we limit our baseline method to 1-shot prompting to assess its effectiveness in enhancing the model's performance on specific downstream tasks.

# 5 EXPERIMENTAL ANALYSIS

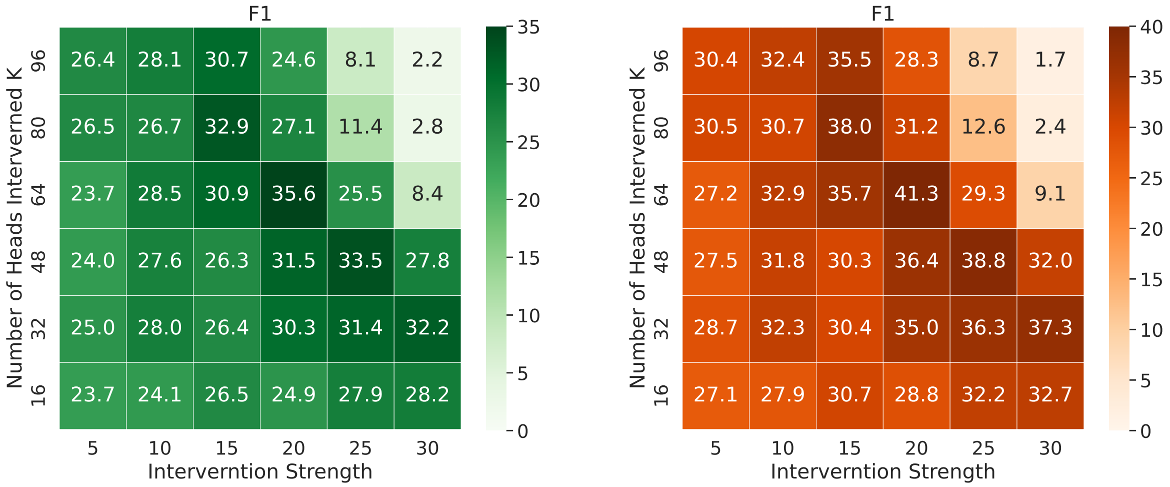
In experimental analysis, we evaluate ASMR from three different perspectives: Performance of ASMR on CLIR benchmarks, Probing accuracy analysis, and hyperparameter optimization for ASMR.

## 5.1 Performance of ASMR on CLIR benchmarks

we delve into the performance of ASMR and compare it to existing methods. Our analysis covers both the MKQA and XOR-TyDi QA datasets to provide a thorough evaluation of our approach.

***ASMR performance on MKQA Dataset.*** ASMR exhibits a significant improvement in the MKQA dataset as shown in Table 1, particularly when benchmarked against prior CLIR methods and other LLM adaptation techniques. For instance, the Sentri model, a strong performer among previous CLIR methods, shows an average R@2kt score of 54.6 across all languages. ASMR, on the other hand, achieves an impressive average score of 55.4, marking a noticeable improvement. This enhancement is even more pronounced in certain languages. In the German language, ASMR outperforms Sentri by 2.6 points, moving from a score of 56.5 to 59.0. Additionally, when compared to LLM adaptation methods like few-shot prompting, which scores an average of 46.1, ASMR's average score reflects a substantial 9.3-point increase. Such improvements underscore the efficacy of ASMR in leveraging LLMs for better CLIR.

***ASMR performance on XOR-TyDi QA Dataset.*** From Table 2, we can tell that the ASMR method's performance on the XOR-TyDi QA dataset underscores its effectiveness in cross-lingual question answering. When compared with existing CLIR methods, ASMR demonstrates a robust enhancement in F1 scores. The BM25 + Ext.reader + MT method, for instance, shows an F1 score of 18.9 for the Bengali language, while ASMR notably improves on this, achieving an F1 score of 36.4. This marks an impressive 17.5-point increase, showcasing ASMR's superior handling of low-resource languages. Moreover, against the backdrop of LLM adaptation methods, ASMR continues to shine. Taking the BLOOM-7B model with supervised fine-tuning as a reference, which scores 44.5 in the Arabic language, ASMR registers a remarkable 7.7-point jump to 55.2. This consistent outperformance across different languages affirms



**(a) Varying hyparameters for BLOOM**

**(b) Varying hyparameters for Llama2-7B**

**Figure 3: Results on XOR-TyDi QA datasets with varying hyperparameters for different LLMs: BLOOM and Llama2-7B. Metrics have been averaged over 5 random seeds.**

the advanced capability of ASMR to accurately understand and respond to multilingual queries, setting a new standard for CLIR systems.

***Effectiveness of Steering Vectors Across Language Families.*** Evaluating ASMR's cross-language performance, it becomes evident that the method has a broad impact across language families. Looking at Table 1, ASMR shows a marked improvement in languages like Turkish, where it achieves a score of 56.3 on the MKQA dataset, outperforming the Sentri model by 3.3 points. Similarly, in Table 2, ASMR demonstrates its strength in low-resource languages such as Finnish, with an F1 score increase from 45.5 to 53.9 when compared to Sentri. These results highlight ASMR's adaptability and accuracy in handling diverse linguistic structures, from agglutinative languages like Turkish to morphologically rich languages like Finnish. The method's capacity to enhance retrieval quality in both well-represented and underrepresented languages showcases its potential to democratize information access across language barriers effectively.

***ASMR's Consistent Improvement Across Different LLMs.*** The ASMR approach demonstrates remarkable consistency in its performance across different LLMs, as evidenced by the results on both the MKQA and XOR-TyDi QA datasets. For the BLOOM-7B model, ASMR boosts the average score to 55.4 on the MKQA dataset and to 35.1 on the F1 metric for the XOR-TyDi QA dataset, surpassing both the supervised fine-tuning and few-shot prompting methods. This trend is mirrored in the performance of the LLama2-7B model, where ASMR again leads to the highest average scores of 55.4 and 31.0 respectively. Such uniformity in performance, regardless of the underlying LLM architecture, not only attests to the robustness of the ASMR method but also suggests its scalability. By consistently enhancing retrieval and question-answering capabilities across diverse language datasets, ASMR proves to be a versatile and reliable technique for CLIR tasks, independent of the specific LLM employed.

## 5.2 Probing Accuracy Analysis

Our experiments include an analysis of probing classifier accuracies for two critical attributes, "Accuracy in Content" and "Coherence in Language," using the Llama2-7B model. The results are visualized in heatmaps (Figure 2) that display the performance of different attention heads across the model's layers.

***Accuracy in Content.*** As shown in Figure 2a, the middle layers of the model demonstrate higher classifier accuracies for content accuracy. This indicates their key role in understanding and accurately processing content. The standout performance of the 9th attention head in the 14th layer, with an accuracy of 79.3%, suggests its specialization in critical aspects of content analysis.

***Coherence in Language.*** Contrastingly, Figure 2b illustrates that both the lower and upper layers of the model excel in ensuring language coherence. The lower layers primarily handle syntactic structures, while the upper layers refine these into coherent outputs. The peak accuracy in language coherence, 98.5%, is observed in the 13th head of the 28th layer.

***Layer-Specific Functions in LLMs.*** The combined findings from these heatmaps reveal a layered functional architecture within the LLM. Middle layers are crucial for content comprehension and processing, whereas lower and upper layers collectively maintain language coherence. This differentiation in layer functionality underlines the model's sophisticated approach to handling the complexities of multilingual information processing.

## 5.3 Hyperparameter Optimization

We focused on optimizing two key hyperparameters: the number of heads intervened upon ($K$) and the intervention strength ($\alpha$),

**Table 3: Comparisons between AMSR and baseline method on two cases, one in French, the other in German. These cases show that ASMR is capable of generating more detailed and linguistically precise responses.**

| | **Case 1 (German)** |
|---|---|
| Query | Was sind die neuesten Durchbrüche in der Krebsforschung?<br>(English: What are the latest breakthroughs in cancer research?) |
| Baseline | Die neuesten Durchbrüche umfassen Fortschritte in der zielgerichteten Therapie und Immuntherapie.<br>(English: Recent breakthroughs include advancements in targeted therapy and immunotherapy.) |
| AMSR | Zu den jüngsten Fortschritten in der Krebsforschung gehören die Entwicklung neuer zielgerichteter Therapien, die Verbesserung der Immuntherapie, die Erforschung von Krebsgenomik und der Einsatz künstlicher Intelligenz für personalisierte Behandlungspläne.<br>(English:Recent advances in cancer research include the development of new targeted therapies, improvements in immunotherapy, research into cancer genomics, and the use of artificial intelligence for personalized treatment plans.) |
| | **Case 2 (French)** |
| Query | Quelles sont les causes du changement climatique?<br>(English: What are the causes of climate change?) |
| Baseline | Le changement climatique est principalement causé par les émissions de carbone et la déforestation.<br>(English: Climate change is primarily caused by carbon emissions and deforestation.) |
| AMSR | Les principales causes du changement climatique incluent non seulement les émissions de gaz à effet de serre comme le CO2, mais également la déforestation, l'urbanisation croissante, et l'agriculture intensive.<br>(English: The main causes of climate change include greenhouse gas emissions like CO2, as well as deforestation, increasing urbanization, and intensive agriculture.) |

to determine their impact on the F1 score for the XOR-TyDi QA dataset.

***Optimization for BLOOM.*** For the BLOOM model, we found that optimal performance was achieved with $K = 64$ and $\alpha = 20$, indicating a balanced approach in steering vector intervention. This specific combination of $K$ and $\alpha$ maximizes the model's F1 score, suggesting it effectively balances the influence on attention heads without overfitting or introducing bias.

***Optimization for Llama2-7B.*** Similar to BLOOM, the Llama2-7B model also reached its best F1 score at $K = 64$ and $\alpha = 20$. This consistency across models implies that both BLOOM and Llama2-7B function optimally within these hyperparameter settings, striking a balance between model influence and generalization capability.

***Consistency and Over-Intervention Risks.*** The identical optimal settings for both BLOOM and Llama2-7B ($K = 64$ and $\alpha = 20$) highlight the robustness and broad applicability of our steering vector approach. However, increasing $\alpha$ beyond 20 resulted in reduced F1 scores for both models, warning against the risks of excessive intervention. Over-steering can adversely affect the model's flexibility and accuracy, underscoring the importance of a measured approach in applying steering vectors for optimal performance across languages and tasks.

### 5.4 Case Study

We show two cases in Table 3 to better evaluate the performance of AMSR. In the first case, ASMR's response expands on the baseline model's answer by specifying "greenhouse gas emissions" and adding other significant factors like urbanization and intensive agriculture. This added detail provides a more comprehensive understanding of the issue, demonstrating ASMR's ability to deliver a richer and more informative answer. For the second case, ASMR delivers a response that goes beyond the baseline model's general statement, detailing additional advancements like emotional recognition and human-robot interaction. This illustrates the ASMR model's capacity to provide a deeper, more nuanced exploration of the topic. The two cases demonstrate ASMR's proficiency in generating responses that are both linguistically accurate and content-rich, offering a clear advancement over traditional post-hoc translation methods in CLIR.

## 6 CONCLUSION

In this paper, we have presented a novel method in CLIR by integrating LLMs with innovative 'steering vectors.' Our method, ASMR, has demonstrated exceptional performance in processing and understanding multilingual queries, surpassing existing benchmarks on XOR-TyDi QA, and MKQA datasets. This success is attributed to the precise tuning of LLMs, through steering vectors that effectively guide the model in accurately interpreting content and maintaining language coherence. We will continue to refine this approach to revolutionize how we access and interact with multilingual information in the digital age.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.U2336202) and the National Natural Science Foundation of China (Grant No.U21B2009).

## REFERENCES

[1] Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. https://openreview.net/forum?id=ryF7rTqgl

[2] Alon Albalak, Sharon Levy, and William Yang Wang. 2023. Addressing Issues of Cross-Linguality in Open-Retrieval Question Answering Systems For Emergent Domains. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Dubrovnik, Croatia, 1–10. https://doi.org/10.18653/v1/2023.eacl-demo.1

[3] Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual Open-Retrieval Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 547–564. https://doi.org/10.18653/v1/2021.naacl-main.46

[4] Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H. Clark, and Eunsol Choi. 2022. MIA 2022 Shared Task: Evaluating Cross-lingual Open-Retrieval Question Answering for 16 Diverse Languages. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*. Association for Computational Linguistics, Seattle, USA, 108–120. https://doi.org/10.18653/v1/2022.mia-1.11

[5] Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. arXiv:2107.11976 [cs.CL]

[6] Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. Generating Synthetic Documents for Cross-Encoder Re-Rankers: A Comparative Study of ChatGPT and Human Experts. arXiv:2305.02320 [cs.IR]

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL]

[8] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (March 2022), 207–219. https://doi.org/10.1162/coli_a_00422

[9] Davis Brown, Charles Godfrey, Cody Nizinski, Jonathan Tu, and Henry Kvinge. 2023. Robustness of edited neural networks. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*. https://openreview.net/forum?id=JAjH6VANZ4

[10] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827 [cs.CL]

[11] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. CorpusBrain: Pre-Train a Generative Retrieval Model for Knowledge-Intensive Language Tasks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 191–200. https://doi.org/10.1145/3511808.3557271

[12] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1edEyBKDS

[13] Mikhail Fain, Niall Twomey, and Danushka Bollegala. 2021. Backretrieval: An Image-Pivoted Evaluation Metric for Cross-Lingual Text Representations Without Parallel Corpora. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2106–2110. https://doi.org/10.1145/3404835.3463027

[14] Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2023. Knowledge Refinement via Interaction Between Search Engines and Large Language Models. arXiv:2305.07402 [cs.CL]

[15] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL]

[16] Taicheng Guo, Lu Yu, Basem Shihada, and Xiangliang Zhang. 2023. Few-Shot News Recommendation via Cross-Lingual Transfer. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 1130–1140. https://doi.org/10.1145/3543507.3583383

[17] Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5427–5444. https://doi.org/10.18653/v1/2020.emnlp-main.438

[18] Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=EldbUlZtbd

[19] Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and Editing Knowledge Representations in Language Models. arXiv:2304.00740 [cs.CL]

[20] Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2023. Language Agnostic Multilingual Information Retrieval with Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 9133–9146. https://doi.org/10.18653/v1/2023.findings-acl.581

[21] Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1024–1035. https://aclanthology.org/2022.coling-1.86

[22] Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving Cross-Lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1048–1056. https://doi.org/10.1145/3539597.3570468

[23] Zhiqi Huang, Hansi Zeng, Hamed Zamani, and James Allan. 2023. Soft Prompt Decoding for Multilingual Dense Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1208–1218. https://doi.org/10.1145/3539618.3591769

[24] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=6t0Kwf8-jrj

[25] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[26] Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2023. NeuralMind-UNICAMP at 2022 TREC NeuCLIR: Large Boring Rerankers for Cross-lingual Retrieval. arXiv:2303.16145 [cs.IR]

[27] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* 9 (2021), 962–977. https://doi.org/10.1162/tacl_a_00407

[28] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. European Language Resources Association, Marseille, France, 26–31. https://aclanthology.org/2020.clssts-1.5

[29] Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-Text Multi-View Learning for Passage Re-Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1803–1807. https://doi.org/10.1145/3404835.3463048

[30] 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[31] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative Discriminator Guided Sequence Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 4929–4952. https://doi.org/10.18653/v1/2021.findings-emnlp.424

[32] 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 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[33] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent World Representations: Exploring a Sequence

Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=DeG07_TcZvT

[34] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv:2306.03341 [cs.LG]

[35] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=3s9IrEsjLyk

[36] Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning Cross-Lingual IR from an English Retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 4428–4436. https://doi.org/10.18653/v1/2022.naacl-main.329

[37] Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and Xinyu Zhang. 2023. Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval. arXiv:2304.01019 [cs.IR]

[38] Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2023. Boosting Zero-shot Cross-lingual Retrieval by Training on Artificially Code-Switched Data. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 3096–3108. https://doi.org/10.18653/v1/2023.findings-acl.193

[39] Robert Litschko, Ivan Vulić, and Goran Glavaš. 2022. Parameter-Efficient Neural Reranking for Cross-Lingual and Multilingual Retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1071–1082. https://aclanthology.org/2022.coling-1.90

[40] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 342–358. https://doi.org/10.1007/978-3-030-72113-8_23

[41] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On Cross-Lingual Retrieval with Multilingual Text Encoders. *Inf. Retr.* 25, 2 (jun 2022), 149–183. https://doi.org/10.1007/s10791-022-09406-x

[42] Jiapeng Liu, Xiao Zhang, Dan Goldwasser, and Xiao Wang. 2020. Cross-Lingual Document Retrieval with Smooth Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3616–3629. https://doi.org/10.18653/v1/2020.coling-main.323

[43] Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. https://arxiv.org/pdf/2007.15207.pdf

[44] Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative and Pseudo-Relevant Feedback for Sparse, Dense and Learned Sparse Retrieval. arXiv:2305.07477 [cs.IR]

[45] Iain Mackie, Ivan Sekulic, Shubham Chatterjee, Jeffrey Dalton, and Fabio Crestani. 2023. GRM: Generative Relevance Modeling Using Relevance-Aware Sample Estimation for Document Retrieval. arXiv:2306.09938 [cs.IR]

[46] C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. https://books.google.co.jp/books?id=t1PoSh4uwVcC

[47] Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. arXiv:2303.06573 [cs.IR]

[48] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=-h6WAS6eE4

[49] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing Implicit Assumptions in Text-to-Image Diffusion Models. arXiv:2303.08084 [cs.CV]

[50] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[52] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. arXiv:2306.17563 [cs.IR]

[53] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2018. Learning To Generate Reviews and Discovering Sentiment. https://openreview.net/forum?id=SJ71VXZAZ

[54] Houxing Ren, Linjun Shou, Ning Wu, Ming Gong, and Daxin Jiang. 2022. Empowering Dual-Encoder with Query Generator for Cross-Lingual Dense Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3107–3121. https://doi.org/10.18653/v1/2022.emnlp-main.203

[55] Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large Language Models are Strong Zero-Shot Retriever. arXiv:2304.14233 [cs.CL]

[56] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-Lingual Training of Dense Retrievers for Document Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 251–253. https://doi.org/10.18653/v1/2021.mrl-1.24

[57] Nikita Sorokin, Dmitry Abulkhanov, Irina Piontkovskaya, and Valentin Malykh. 2022. Ask Me Anything in Your Native Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 395–406. https://doi.org/10.18653/v1/2022.naacl-main.30

[58] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Michael Bendersky. 2022. QUILL: Query Intent with Large Language Models using Retrieval Augmentation and Multi-stage Distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Abu Dhabi, UAE, 492–501. https://doi.org/10.18653/v1/2022.emnlp-industry.50

[59] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. arXiv:2205.05124 [cs.CL]

[60] Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual Representation Distillation with Contrastive Learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 1477–1490. https://doi.org/10.18653/v1/2023.eacl-main.108

[61] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 21831–21843. https://proceedings.neurips.cc/paper_files/paper/2022/file/892840a6123b5ec99ebaab8be1530fba-Paper-Conference.pdf

[62] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672 [cs.CL]

[63] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4593–4601. https://doi.org/10.18653/v1/P19-1452

[64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[65] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 10014–10037. https://doi.org/

10.18653/v1/2023.acl-long.557

[66] Zhucheng Tu and Sarguna Janani Padmanabhan. 2022. MIA 2022 Shared Task Submission: Leveraging Entity Representations, Dense-Sparse Hybrids, and Fusion-in-Decoder for Cross-Lingual Question Answering. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*. Association for Computational Linguistics, Seattle, USA, 100–107. https://doi.org/10.18653/v1/2022.mia-1.10

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[68] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. arXiv:2303.07678 [cs.IR]

[69] Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021. Adversarial Domain Adaptation for Cross-Lingual Information Retrieval with Multilingual BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 3498–3502. https://doi.org/10.1145/3459637.3482050

[70] John Wieting, Jonathan Clark, William Cohen, Graham Neubig, and Taylor Berg-Kirkpatrick. 2023. Beyond Contrastive Learning: A Variational Generative Model for Multilingual Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 12044–12066. https://doi.org/10.18653/v1/2023.acl-long.673

[71] BigScience Workshop. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 [cs.CL]

[72] Linlong Xu, Baosong Yang, Xiaoyu Lv, Tianchi Bi, Dayiheng Liu, and Haibo Zhang. 2021. Leveraging Advantages of Interactive and Non-Interactive Models for Vector-Based Cross-Lingual Information Retrieval. arXiv:2111.01992 [cs.CL]

[73] Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A Simple and Effective Method To Eliminate the Self Language Bias in Multilingual Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5825–5832. https://doi.org/10.18653/v1/2021.emnlp-main.470

[74] Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal Sentence Representation Learning with Conditional Masked Language Model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6216–6228. https://doi.org/10.18653/v1/2021.emnlp-main.502

[75] Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Exploiting Neural Query Translation into Cross Lingual Information Retrieval. arXiv:2010.13659 [cs.CL]

[76] 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1637–1640. https://doi.org/10.1145/3397271.3401322

[77] Bryan Zhang and Amita Misra. 2022. Machine translation impact in E-commerce multilingual search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Abu Dhabi, UAE, 99–109. https://doi.org/10.18653/v1/2022.emnlp-industry.8

[78] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 4 (Jun. 2022), 4345–4353. https://doi.org/10.1609/aaai.v36i4.20355

[79] Shunyu Zhang, Yaobo Liang, MING GONG, Daxin Jiang, and Nan Duan. 2023. Modeling Sequential Sentence Relation to Improve Cross-lingual Dense Retrieval. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=-bVsNeR56KS

[80] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023. Towards Best Practices for Training Multilingual Dense Retrieval Models. *ACM Trans. Inf. Syst.* (aug 2023). https://doi.org/10.1145/3613447 Just Accepted.

[81] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2023. Bridging the Gap Between Indexing and Retrieval for Differentiable Search Index with Query Generation. arXiv:2206.10128 [cs.IR]

[82] Shengyao Zhuang, Linjun Shou, and Guido Zuccon. 2023. Augmenting Passage Representations with Query Generation for Enhanced Cross-Lingual Dense Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1827–1832. https://doi.org/10.1145/3539618.3591952

[83] Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large Language Models are Built-in Autoregressive Search Engines. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 2666–2678. https://doi.org/10.18653/v1/2023.findings-acl.167