CLIRudit: Cross-Lingual Information Retrieval of Scientific Documents

Francisco Valentini[1,2] *, **Diego Kozlowski**[2], **Vincent Larivière**[2]

[1]CONICET-Universidad de Buenos Aires.
Instituto de Ciencias de la Computación (ICC). Buenos Aires, Argentina
[2]École de bibliothéconomie et des sciences de l'information.
Université de Montréal. Montréal, Canada
fvalentini@dc.uba.ar, diego.kozlowski@umontreal.ca, vincent.lariviere@umontreal.ca

Abstract

Cross-lingual information retrieval (CLIR) helps users find documents in languages different from their queries. This is especially important in academic search, where key research is often published in non-English languages. We present CLIRudit, a novel English-French academic retrieval dataset built from Érudit, a Canadian publishing platform. Using multilingual metadata, we pair English author-written keywords as queries with non-English abstracts as target documents, a method that can be applied to other languages and repositories. We benchmark various first-stage sparse and dense retrievers, with and without machine translation. We find that dense embeddings without translation perform nearly as well as systems using machine translation, that translating documents is generally more effective than translating queries, and that sparse retrievers with document translation remain competitive while offering greater efficiency. Along with releasing the first English-French academic retrieval dataset, we provide a reproducible benchmarking method to improve access to non-English scholarly content.

1 Introduction

Cross-lingual information retrieval (CLIR) helps users find documents written in languages different from their search queries. This removes the need for proficiency in multiple languages and makes it easier to access valuable information that might otherwise be missed because of language barriers.

CLIR is especially important for academic research. While English is the main language for scientific communication, important work often exists in other languages, particularly in certain fields and historical contexts (Pölönen, 2020; Beigel and Digiampietri, 2022; Khanna et al., 2022). Researchers may overlook key work if they cannot search across languages, especially if they're unfamiliar with technical terms. English is also often used due to the expectation of finding more results, reinforcing bias against documents in other languages.

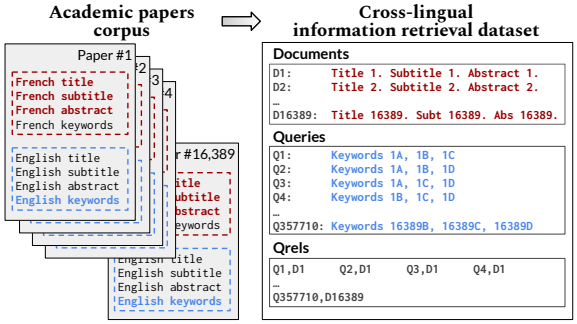

Figure 1: The CLIRudit dataset. We use articles with abstracts and keywords in both French and English. English keywords form the queries, with relevance judged by their presence in each article. Documents consist of the French title, subtitle, and abstract.

Modern information retrieval (IR) systems often use bi-encoder architectures for first-stage retrieval, separately encoding documents and queries as dense embeddings (Devlin et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021). Multilingual extensions of these methods have been effective in general-domain CLIR (Artetxe and Schwenk, 2019; Conneau et al., 2020; Anastasopoulos and Neubig, 2020; Asai et al., 2021a; Nair et al., 2022; Zhang et al., 2023a). Another common approach is to use machine translation (MT) to convert queries or documents to the same language before searching (Galuščáková et al., 2022; Lin et al., 2022; Huang et al., 2023; Lawrie et al., 2024).

Technical texts often use specialized vocabulary and styles that present challenges for MT and multilingual embeddings (Lawrie et al., 2024; Litschko et al., 2025). However, research on CLIR in technical domains is limited (Xu et al., 2016; Zavorin

*Research conducted during a stay at the École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada.

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 226–242
November 8-9, 2025

et al., 2020), and studies focusing specifically on academic content are even scarcer, typically relying on small, curated datasets (Lawrie et al., 2024). As a result, the effectiveness of CLIR methods for academic retrieval remains underexplored.

We address this gap by introducing a dataset for cross-lingual academic search and benchmarking first-stage retrieval methods. Our contributions are:

• A new method for creating academic CLIR datasets using multilingual metadata. We use English keywords as queries and non-English abstracts as documents, allowing evaluation of IR methods on retrieving original-language documents based on author-provided English keywords. This method can be applied to other academic databases and language pairs.

• The release of CLIRudit, a dataset based on Érudit, a Quebec-based non-profit publishing platform (Fig. 1).[1] To our knowledge, this is the first dataset for English-French academic retrieval.

• A thorough empirical comparison of first-stage CLIR methods, including query and document translation, and state-of-the-art dense and sparse retrievers.

• Practical insights to improve the discoverability of non-English scholarly content, which is especially relevant for academic publishing platforms.

Our results show that dense embeddings without translation perform nearly as well as those using MT. Document translation generally improves retrieval more than query translation. While sparse retrievers combined with document translation may not surpass the best dense multilingual methods, they remain competitive and offer advantages in search speed and indexing efficiency.

2 Related work

This section reviews relevant research on academic CLIR, focusing on first-stage retrieval methods, datasets, and bilingual academic corpora.

2.1 Cross-lingual retrieval

Lin et al. (2022) proposed a conceptual framework for CLIR, outlining three main strategies for first-stage retrieval: **document translation** (DT), translating documents into the query language; **query translation** (QT), translating queries into the document language; and **language-independent representations**, encoding queries and documents into a shared vector space for direct retrieval. Since we focus on single-stage retrieval, we do not address later steps of a retrieval pipeline, such as re-ranking or results fusion.

Translation-based methods have been widely used and generally effective, although their success has varied across domains and language pairs. DT combined with neural ranking has shown strong performance in general-domain tasks (Lin et al., 2022; Lawrie et al., 2023b; Lassance et al., 2023), often outperforming QT, which struggles with short, ambiguous queries and limited training data (Galuščáková et al., 2022). However, DT is not a clear winner, with QT performing better in domains like healthcare (Saleh and Pecina, 2020) and in high-resource languages (Huang et al., 2023).

Alternative approaches like probabilistic structured queries (PSQ) generate multiple plausible translations per term using alignment models, offering more flexibility than standard machine translation (Darwish and Oard, 2003; Yang et al., 2024c).

Early studies found a strong link between translation quality and retrieval effectiveness (Zhu and Wang, 2006), but later work found that better MT doesn't always improve retrieval, particularly in specialized domains (Pecina et al., 2014). Recent research suggests a weak positive correlation (Bonifacio et al., 2022) with diminishing returns beyond a certain MT quality level (Zhang and Misra, 2022).

Multilingual bi-encoders avoid MT entirely by using multilingual pretrained models (Jiang et al., 2020; Bonifacio et al., 2022; Nair et al., 2022, 2023). These methods can reduce indexing costs but often perform worse than MT-based retrieval, with QT or DT followed by monolingual retrieval frequently achieving better first-stage results (Litschko et al., 2019; Asai et al., 2021a; Lin et al., 2022; Nair et al., 2023; Lawrie et al., 2023b).

Recent methods like translate-train (Nair et al., 2022) and translate-distill (Yang et al., 2024b) integrate MT into training, allowing bi-encoders to jointly learn retrieval and translation; unlike translate-test methods like DT and QT, which translate only at test-time. Translate-distill further uses distillation from cross-encoders, achieving strong results across multiple languages. Additionally, large decoder-only language models (LLMs) have been adapted as bi-encoders for dense retrieval (Lee et al., 2024; Li et al., 2025).

[1] https://hf.co/datasets/ftvalentini/clirudit

2.2 CLIR datasets

Well-documented and diverse datasets are crucial for advancing CLIR because they enable training and evaluation across languages and domains.

Shared evaluation initiatives like TREC (Voorhees, 2005) and CLEF (Chen, 2002) provide manually curated test collections with human-generated queries and relevance judgments gathered by pooling top-ranked results. NeuCLIR (TREC 2022) focuses on neural CLIR, alongside other datasets such as BETTER (Soboroff, 2023) and HC4 (Lawrie et al., 2022). While these collections are usually carefully designed, they are typically small, often with fewer than 1,000 queries. Galuščáková et al. (2022) provide a comprehensive survey of such resources.

Sentence-level retrieval datasets are also common, such as BUCC, Tatoeba (Siddhant et al., 2020), and STS17/STS22 (Cer et al., 2017; Chen et al., 2022), which focus on matching similar sentences across languages.

To address scale limitations, recent work has explored automatic dataset creation. For example, Mayfield et al. (2023) used LLMs to generate English queries from target-language documents. Wikipedia's multilingual, structured content has also been used for automatic dataset creation, as seen in MuSeCLIR (Li et al., 2022), MKQA (Longpre et al., 2021), WikiCLIR (Sasaki et al., 2018), CLIRMatrix (Sun and Duh, 2020), and AfriCLIRMatrix (Ogundepo et al., 2022).

2.3 Academic datasets

Some prior datasets address CLIR in technical domains. For example, Xu et al. (2016) study cross-language technical question retrieval, CLEF eHealth simulates medical search by non-experts (Galuščáková et al., 2022), and MATERIAL covers law, security, and health topics (Zavorin et al., 2020). A close reference to our work is NeuCLIR 2023's technical track, which contains 40 English queries to retrieve Chinese academic abstracts across Chemistry, Economics, Physics, Biology, and Medicine (Lawrie et al., 2024). NeuCLIR 2024 also featured a technical task but their proceedings were unavailable at the time of writing.

Beyond CLIR-specific datasets, some parallel academic corpora similar to the one we use include academic metadata aligned across languages. SciPar (Roussis et al., 2022) compiles bilingual titles and abstracts from theses and dissertations. Other examples mentioned in Roussis et al. (2022) include SciELO (Neves et al., 2016, English, Portuguese, Spanish), ASPEC (Nakazawa et al., 2016, English, Japanese, Chinese), CAPES (Soares et al., 2018, Brazilian academic works), and EDP (Névéol et al., 2018, English-French biomedical texts). In the biomedical domain, MEDLINE (Wu et al., 2011) and BVS (Soares and Krallinger, 2019) provide multilingual aligned abstracts. Niu and Jiang (2024) introduce a dataset of translated abstracts from journals in translation studies.

These corpora mainly support MT by providing parallel abstracts and titles, often with aligned sentences. Our work differs by using keywords as queries of a CLIR dataset. Among existing corpora, only CAPES and BVS include multilingual keywords suitable for this task, but they are not publicly available at the time of writing.

3 Evaluation data

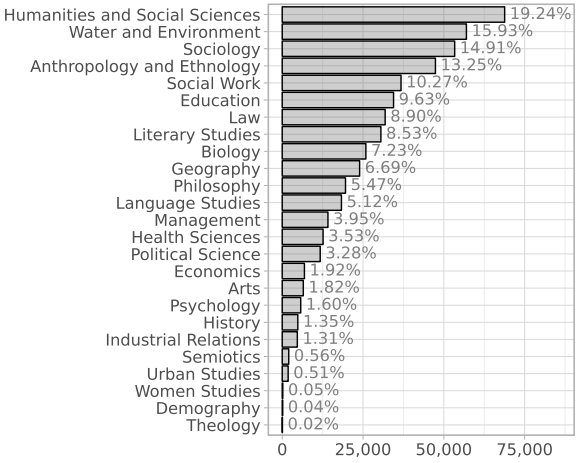

Figure 2: Number of queries per disciplines in the CLIRudit dataset. A query inherits the disciplines of the articles containing its keywords. Since queries can originate from multiple articles and articles can have multiple disciplines, percentages do not sum to 100%.

To evaluate academic CLIR methods, we built CLIRudit using data from Érudit[2], a Quebec-based Canadian platform that publishes research in the arts, humanities, and social sciences. Érudit's journals are selected by a scientific committee and meet national quality standards, ensuring the relevance and quality of the content.

We focused exclusively on research articles that included both English and French abstracts and keywords, provided by the authors. From each

[2]https://www.erudit.org/en/

article's metadata, we extracted the title, subtitle, abstract, and keywords.

Following the standard CLIR task setup, with English queries targeting non-English documents (Lawrie et al., 2023a, 2024), we built the dataset as follows (see Fig. 1 for an overview):

• **Queries.** Created by combining all possible groups of three English keywords from each article; e.g., an article with keywords {A, B, C, D} generates the queries: "A, B, C", "A, B, D", "A, C, D", and "B, C, D".

• **Relevance judgments.** A document was marked relevant to a query if its English keyword metadata included all three query keywords. This is based on the assumption that authors try to make their work discoverable via those terms.

• **Document collection.** Each document or retrieval unit was built as the concatenation of its French title, subtitle, and abstract.

We chose three-keyword combinations for queries based on preliminary observations. Using only two keywords produced overly broad queries which could apply to many documents even if those specific terms weren't used by the authors; e.g., "*family dynamics, gender identity*" or "*canada, québec*". In contrast, using more than three keywords led to overly narrow queries that were unlikely to reflect realistic user search behavior.

The final dataset contains 357,710 queries derived from 41,594 unique English keywords, with an average query length of 4.8 words (SD = 1.7); and 16,389 French documents from 124 journals across 25 disciplines, with an average document length of 176.7 words (SD = 82.4). Because of the way the dataset was built, all documents in the collection are relevant to at least one query. 99.3% of queries have only one relevant document, showing that most three-keyword combinations are unique to a single article, which highlights the specificity of the queries.

84.9% of the abstracts in the dataset come from articles whose primary language is French, 14.3% from English, and 0.9% from other languages. The most frequent disciplines in the queries are Humanities and Social Sciences, Water and Environment, Sociology, and Anthropology and Ethnology (full distribution in Fig. 2).

CLIRudit simulates a scenario where users know only the relevant terms in English, while the pertinent documents are only in French, with no direct translations available. Our pipeline offers a reproducible method to build CLIR datasets for academic search. Rather than relying on complex heuristics, it leverages the inherent bilingual structure of scientific publications. While this work focuses on English-French retrieval, the method can be extended to other databases and language pairs, facilitating research in cross-lingual scientific retrieval.

4 Models and methods

This section describes the retrieval and MT methods, and evaluation metrics used for benchmarking.

4.1 Retrievers

We tested lexical, sparse, and dense first-stage retrievers, all operating as bi-encoders, encoding queries and documents separately. Due to our relatively small document collection, we used exhaustive nearest-neighbor search. We prioritized well-documented, open-source models.

Dense multilingual retrievers. We evaluated three state-of-the-art bi-encoders for direct CLIR without translation, as they are pretrained and fine-tuned on multilingual data: **mE5**[3] (Wang et al., 2024), **mGTE-dense**[4] (Zhang et al., 2024), and **BGE-m-gemma2**[5] (Li et al., 2025). While mGTE-dense and BGE-m-gemma2 are fine-tuned on some cross-lingual tasks involving mixed-language inputs, mE5 is trained on multilingual but not explicitly cross-lingual data, which may affect CLIR performance.

Dense English retrievers. We included English-focused models to assess two approaches: (1) retrieving French documents translated to English, or (2) leveraging cross-lingual transfer, where models, fine-tuned mainly on one language, perform well on other languages for the same task (Artetxe and Schwenk, 2019; Asai et al., 2021b; Zhang et al., 2023a). We assessed two top English MTEB (Muennighoff et al., 2023) performers as of early 2025: **NV-Embed-v2**[6] (Lee et al., 2024), and **BGE-EN-ICL**[7] (Li et al., 2025). Though targeting English, these models have some multilingual fine-tuning (including French), and their Mistral-7B backbone (Jiang et al., 2023) may also have had

[3] intfloat/multilingual-e5-large
[4] Alibaba-NLP/gte-multilingual-base
[5] BAAI/bge-multilingual-gemma2
[6] nvidia/NV-Embed-v2
[7] BAAI/bge-en-icl

multilingual pretraining. but this information is not publicly available.

French-specialized dense retrievers. Few dense retrievers specialize in non-English languages, and those that do are developed by open source communities and lack thorough documentation. We considered these top performers on the MTEB French benchmark (Ciancone et al., 2024): **Croissant**[8] (from CroissantLLM, Faysse et al., 2024), **Solon**[9], and **Lajavaness**[10], all of which are bilingual at some degree as they include English data in pre-training or fine-tuning.

Dense multi-vector retrievers. ColBERT-style models encode queries and documents into token-level embeddings, enabling fine-grained late interaction and pre-computation of document representations, with strong performance in out-of-domain retrieval (Khattab and Zaharia, 2020; Santhanam et al., 2022b). PLAID (Santhanam et al., 2022a) improves speed using clustering and centroid-based interaction. We tested **PLAID-X**[11] (Yang et al., 2024a), a multilingual ColBERT variant trained via translate-distill, distilling signals from an English cross-encoder and translated passages. It uses multilingual batching to support English queries and French, German, and Spanish documents.

Sparse retrievers. These encode queries and documents as term-weighted vectors, enabling efficient retrieval with inverted indexes (Formal et al., 2022). We tested **BM25** (Robertson et al., 2009), a strong exact-match baseline (Thakur et al., 2021), used on inputs translated into a common language.

Learned sparse models improve retrieval by expanding terms through supervised training (Lin et al., 2022). We assessed **SPLADE++**[12] (monolingual, requires MT into English); and the multilingual **mGTE-sparse** (Zhang et al., 2024) and **BGE-M3-sparse** (Chen et al., 2024), which allow cross-lingual retrieval but lack term expansion, limiting performance when queries and documents share few tokens. We excluded BLADE (Nair et al., 2023), a cross-lingual SPLADE variant with term expansion, due to the lack of an English-French version. Additionally, BLADE has demonstrated lower effectiveness compared to PLAID-X, which we included in our evaluation.

Finally, we tested **PSQ** (Yang et al., 2024c), which enables sparse CLIR without conventional MT by indexing documents in query language tokens using a probabilistic alignment matrix (Yang et al., 2024b).

See Appendix A for further details on the models and their implementations.

4.2 Machine translation

We tested three machine translation models:

- **GPT-4o-mini**[13]. Recent work shows LLMs perform well on document-level MT (Kocmi et al., 2023; Zhang et al., 2023b; Pang et al., 2025). We used a cost-efficient proprietary model which performed competitively on high-resource language pairs (Hendy et al., 2023; Zhu et al., 2024).
- **Llama-3.2**. We used the 3.2B-parameter version as an open-source LLM alternative to GPT, with strong zero-shot capabilities in French to English translation (Zhang et al., 2023b). Open-source models can be advantageous for cost-efficiency and for the ability to fine-tune on domain-specific data.
- **OpusMT**, a 75M-parameter French-English MarianMT encoder-decoder model (Tiedemann et al., 2023) trained on Opus parallel data[14]. While designed for sentence-level MT, we applied it at the document level following Cui et al. (2024). It supports up to 512 tokens, far fewer than the 100k+ limits of GPT and Llama.

For LLM translation we used a zero-shot prompt suited for instruction-tuned LLMs (details in Appendix B). We did not test other strong proprietary translators due to lack of cost-efficient APIs.

Finally, as **gold standard** translations, we used the English translations of the French titles, subtitles, and abstracts provided by the article authors. These reflect the potential performance of each retrieval method using human translations. We did not use the actual French keywords as "gold standard" queries since they do not map one-to-one to the English keywords; using them would alter the original set of evaluation queries and introduce noise into the analysis.

4.3 Evaluation metrics

To measure retrieval performance, we use Recall@100 and Mean Average Precision with a 1000 cutoff rank (MAP), which have been widely used (Nair et al., 2023; Lawrie et al., 2024; Yang et al., 2024c). Whereas Recall@100 is useful to assess

[8] manu/sentence_croissant_alpha_v0.3
[9] OrdalieTech/Solon-embeddings-large-0.1
[10] Lajavaness/bilingual-embedding-large
[11] plaidx-large-clef-mtd-mix-passages-mt5xxl-engeng
[12] naver/splade-cocondenser-ensembledistil

[13] gpt-4o-mini
[14] Helsinki-NLP/opus-mt-fr-en

the effectiveness of methods when used as first-stage retrievers, MAP is more appropriate for measuring overall performance of a method used as a single-stage system (Yang et al., 2024c). We compute 95% bootstrap confidence intervals with 1,000 resamples to assess statistical significance.

To evaluate document translation quality, we used three metrics used in recent works (Sun et al., 2022; Zhang et al., 2022; Zhuocheng et al., 2023): BLONDE (Jiang et al., 2022), document-BLEU (d-BLEU, Liu et al., 2020), and document-chrF (d-chrF, Zhuocheng et al., 2023).

5 Results and analysis

This section analyzes the performance of retrieval and translation models on CLIRudit (Table 1). Due to the large sample size, no confidence interval width exceeded 0.003. Intervals are omitted here for readability (see Appendix C). To account for input length effects, we evaluated each method both at its native input limit and with the same 512-token limit. Results differed by no more than 0.005 from the reported values, small enough to not affect general trends.

We now discuss key findings from the results.

1. Without translation, dense retrievers excel, even without multilingual retrieval fine-tuning. NV-Embed-v2 and PLAID-X achieved the highest MAP, while BGE-m-gemma2 led in Recall@100. Interestingly, NV-Embed-v2 is not reported to have multilingual capabilities; though its fine-tuning data, which is English-only for retrieval, includes French in STS17 and STS22 sentence pairs (Cer et al., 2017; Chen et al., 2022).

BM25 with the French analyzer performed poorly without MT due to the query-document language mismatch, but still had non-zero results. This shows CLIRudit has some query-document lexical overlap; manual inspection revealed shared terms like proper nouns, Latin terms, and acronyms.

Among sparse models, SPLADE++ outperformed mGTE-sparse and BGE-M3-sparse, likely thanks to query expansion mitigating the language mismatch. PSQ addresses this mismatch via probabilistic translation, reaching MAP comparable to larger dense models like mE5 and mGTE-dense.

2. Document translation can improve dense retrievers. DT with GPT-4o-mini improved dense retriever MAP by up to 10% and Recall@100 by up to 5% (Fig. 3, left). The highest MAP overall came from NV-Embed-v2+DT and PLAID-X+DT with GPT-4o-mini. However, translation sometimes hurt performance, especially with QT, affecting models like mE5, BGE-EN-ICL, and even top-performing ones like NV-Embed-v2 and PLAID-X.

Manual review showed QT can reduce recall by mistranslating proper nouns with identical cross-language spelling. For example, "*Goose Bay*" (a Canadian town) was incorrectly translated as "*Baie aux Oies*" instead of remaining unchanged.

3. Document translation usually outperformed query translation for sparse retrievers. Translation had a modest effect on dense models but significantly boosted sparse retrieval. Moreover, DT consistently outperformed QT (Fig. 3 right), especially for BM25 and SPLADE++, with SPLADE++ plus DT nearing the top dense retriever MAP, and also outperforming the PSQ probabilistic translation method (Table 1).

While DT may offer richer context than QT (Galuščáková et al., 2022; Lin et al., 2022), DT outperforming QT is expected for SPLADE++ since it's trained only in English. In contrast, mGTE-sparse and BGE-M3-sparse performed similarly with QT and DT.

Manual inspection of BM25 cases where DT outperformed QT shows that DT can preserve key terms better. For example, "*fair innings*" correctly remains unchanged with DT to English, but translating the query to French yields "*juste part*", which isn't in the original document. Similarly, the term "*beck*" in a query about the surname of a social scientist is correctly preserved in DT, but mistranslated as "*appel*" in the query (French for "call"), making the document irretrievable.

4. Document translation quality correlated with retrieval performance. GPT-4o-mini led in document translation quality (BLEU=34.41, BLONDE=49.32, chrF=63.83), followed closely by Llama (BLEU=31.27, BLONDE=46.52, chrF=61.56), with OpusMT trailing far behind (BLEU: 10.77, BLONDE: 19.35, chrF: 36.15). This ranking mirrors their retrieval performance, where GPT-4o-mini systematically outperformed Llama, which in turn outperformed OpusMT (Table 1). While these results indicate a correlation between translation and retrieval quality, quantifying MT's exact contribution requires further study beyond the scope of this paper.

5. Top dense retrievers approached gold translation recall. Models like NV-Embed-v2, BGE-m-gemma2, BGE-EN-ICL, and PLAID-X, performed close to their gold translation recall (Fig.

Machine Trans. (→) / Retriever (↓)	MAP						Recall@100					
	None	Query (GPT4)	Doc.			Gold	None	Query (GPT4)	Doc.			Gold
			Opus	Llama	GPT4				Opus	Llama	GPT4	
mE5	0.434	0.412	0.448	0.480	0.490	0.526	0.784	0.760	0.790	0.817	0.823	0.840
mGTE-dense	0.450	0.445	0.452	0.459	0.468	0.496	0.820	0.813	0.820	0.834	0.837	0.849
BGE-m-gemma2	0.571	0.543	0.533	0.548	0.560	0.571	**0.903**	**0.895**	**0.894**	**0.908**	**0.910**	**0.917**
NV-Embed-v2	**0.580**	**0.575**	**0.541**	0.569	**0.586**	0.600	0.895	0.889	0.866	0.887	0.892	0.894
BGE-EN-ICL	0.507	0.441	0.411	0.486	0.501	0.535	0.857	0.810	0.760	0.831	0.837	0.861
Croissant	0.358	0.365	0.325	0.345	0.357	0.376	0.793	0.794	0.748	0.773	0.781	0.794
Solon	0.507	0.516	0.502	0.520	0.536	0.555	0.856	0.858	0.845	0.860	0.866	0.870
Lajavaness	0.472	0.454	0.431	0.457	0.470	0.486	0.848	0.838	0.817	0.836	0.843	0.849
PLAID-X	0.578	0.548	0.539	**0.572**	**0.586**	0.605	0.870	0.854	0.845	0.869	0.874	0.879
SPLADE++	0.284	0.426	0.530	0.548	0.572	0.609	0.604	0.753	0.836	0.853	0.864	0.875
mGTE-sparse	0.169	0.434	0.401	0.405	0.428	0.487	0.443	0.763	0.737	0.760	0.771	0.805
BGE-M3-sparse	0.177	0.458	0.413	0.434	0.460	0.511	0.449	0.781	0.738	0.763	0.778	0.807
BM25	0.181	0.390	0.488	0.513	0.549	**0.611**	0.417	0.706	0.789	0.815	0.832	0.861
PSQ	0.440	-	-	-	-	-	0.756	-	-	-	-	-

Table 1: MAP and Recall@100 in CLIRudit. Best column scores are in bold; best row scores per metric are underlined, excluding gold translation. Statistical significance is shown in Appendix C for better readability.

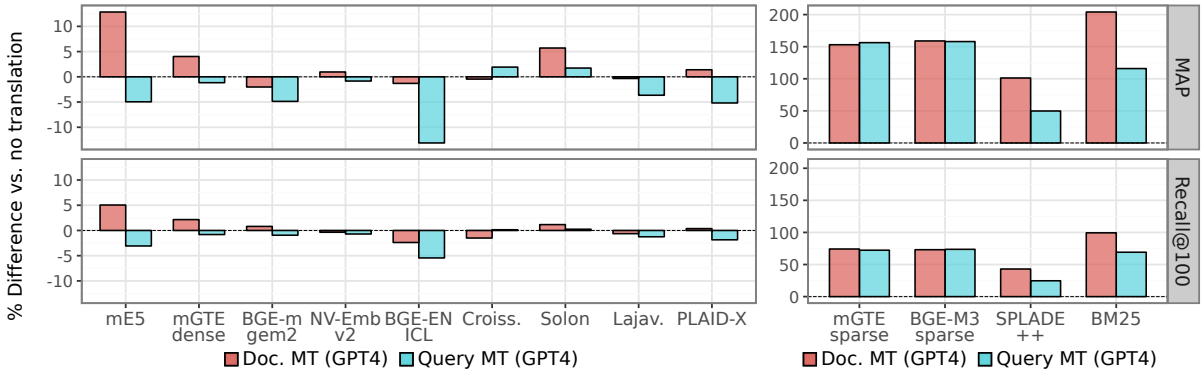

Figure 3: % difference in MAP and Recall@100 of document translation (red) and query translation (blue) compared to no translation. Positive (negative) values indicate improvement (degradation) with translation. For ease of visualization, sparse models are shown with a different scale and only GPT translation is considered.

4). Except for BGE-m-gemma2, gaps in MAP were larger, indicating potential for better ranking.

Sparse models BM25 and SPLADE++ achieved the highest MAP with gold translations (Table 1), highlighting the impact of translation quality. Because CLIRudit queries are keywords and documents are abstracts, sparse models naturally perform well with accurate translations. SPLADE's smaller gap to gold as compared to other sparse methods suggests greater robustness to translation errors, likely due to query expansion.

6. Performance varies significantly across disciplines. Considering the best-performing approach for each retriever, MAP was on average higher in Industrial Relations, Theology, Women's Studies, Psychology, Management, and Economics, and lower in Philosophy and Law (Fig. 5). While Croissant was typically the weakest across disciplines, no translation-retriever combination consistently outperformed the others.

6 Discussion

Dense single-vector retrievers based on large decoder-only models (e.g., NV-Embed-v2, BGE-m-gemma2) achieve near gold translation-augmented performance without additional training, which may result from pretraining on large corpora and cross-lingual transfer capabilities. A smaller, CLIR-specialized model, PLAID-X, also performed competitively; at the expense of needing language- and task-specific training data and having higher search latency due to its multi-vector design (Santhanam et al., 2022a). Both dense approaches avoid the overhead of translating the entire corpus, but large models may incur high index-

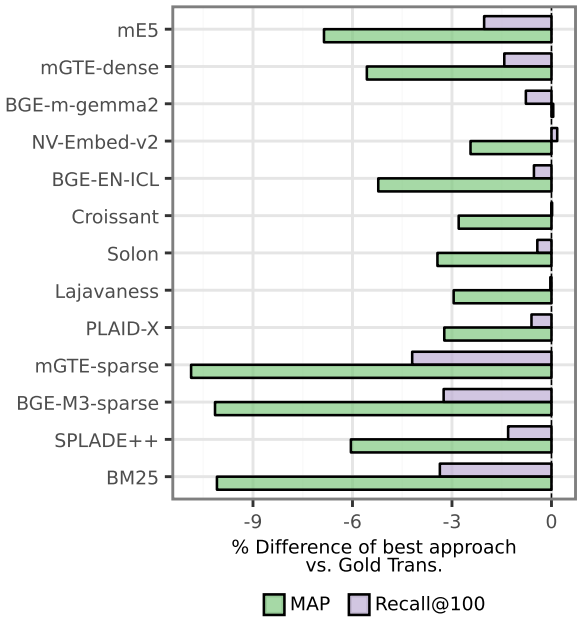

Figure 4: % difference in MAP (green) and Recall@100 (purple) for the best-performing approach of each retriever, relative to gold-standard translations.

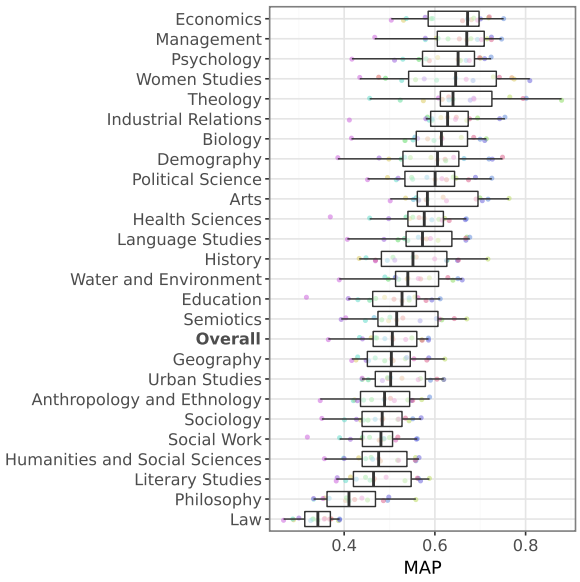

Figure 5: MAP of retrievers across CLIRudit disciplines. Each dot represents a method's MAP in a discipline's queries, using its best translation method (excluding gold). Dot colors indicate retrievers: Croissant (pink) often performs worst, while the best varies by discipline.

ing costs on large collections.

Sparse retrievers, lexical or learned, offer faster indexing and search, but need translation to narrow the gap with dense methods, and still fall short in overall performance. DT outperformed QT, likely because it provides richer context; and it can be done offline, which is important when using costly MT systems. QT enables quicker experimentation by avoiding corpus reindexing with each new model, but usually with lower accuracy (Lin et al., 2022; Galuščáková et al., 2022). Ultimately, the choice of method comes down to balancing retrieval performance, indexing and search latency, and translation costs.

Our dataset uses keyword-based queries, reflecting how authors describe their work to make it discoverable. This assumes users know the right keywords, shifting the dataset challenge to language differences in a technical domain rather than query formulation. This allows meaningful analysis, though it's unclear how system rankings might change with other types of queries, e.g., natural language questions. Our approach aligns with other datasets using non-natural or generated queries, such as SCIDOCS, DBPedia (Thakur et al., 2021), WikiCLIR (Sasaki et al., 2018), and CLIRMatrix (Sun and Duh, 2020).

Like all IR datasets, ours has limitations in scope and collection method, so we encourage evaluation on many, diverse datasets. As the first English-French academic retrieval dataset, CLIRudit adds to this diversity and complements existing resources.

7 Conclusions

We introduced a method for building CLIR datasets from bilingual metadata in scientific publications. By using keywords as queries and abstracts as documents, this approach enables automated, scalable creation of large evaluation resources without manual annotation or complex heuristics. We applied it to produce CLIRudit, the first English-French CLIR dataset for academic search, based on a real-world database.

Evaluations of single-stage methods on CLIRudit showed that: (1) state-of-the-art dense bi-encoders achieved strong cross-lingual performance without translation, nearing monolingual retrieval with gold translations; (2) sparse retrievers with document translation were competitive; and (3) document translation generally outperformed query translation, likely due to richer context.

These results have practical implications for academic search systems. Large dense retrievers deliver the best performance, but the strong results of sparse retrievers with document translation suggest a viable alternative that may be more practical to implement at scale. This is particularly relevant

for academic publishing platforms like Érudit that aim to make their content more discoverable to researchers.

Our method can be applied to other academic databases and language pairs, supporting broader research in cross-lingual access to scientific knowledge.

Limitations

Our dataset's document collection includes only relevant documents, unlike in real applications where relevant documents might coexist with a much larger collection. The values reported may not be representative of real-world settings. The reported metrics should be used to compare methods rather than to provide absolute performance estimates, which is standard practice in IR research (Thakur et al., 2021).

Our dataset may also contain some false negatives: some relevant documents may not be labeled as such if some authors did not include some suitable keywords in the metadata, while others did. However, because queries consist of three keywords, they are relatively specific, likely reducing false negatives, as it is unlikely that there is more than one document in the collection relevant to a narrow query.

We found that the proprietary GPT-4o-mini LLM outperformed the open-source Llama 3.2 and the smaller OpusMT encoder-decoder for zero-shot translation. Further exploration with few-shot prompting or fine-tuning may improve the performance of the open-source models. In addition, OpusMT is not optimized for document translation, so using sentence-level translation may be more optimal. However, this approach requires a more complex pipeline with sentence splitting and risks losing cross-sentence coherence.

Possible data contamination is a concern for fair evaluation: our test set may appear in the training data of pre-trained models, especially LLMs used for translation and retrievers initialized from LLMs, such as NV-Embed-v2 and BGE-m-gemma2. This could lead to inflated results, but is difficult to verify due to the lack of information about the exact training data of these models (Sainz et al., 2023; Oren et al., 2024).

Our dataset is limited to keyword-based queries and metadata-only documents. Results may differ with other query types, e.g. natural language questions, or full-text documents. Future work could explore approaches that use other types of queries or full-text representations. We also focused on French, a high-resource language; performance may vary in low-resource settings due to lower translation quality and limited training data for retrievers.

We tested single-stage retrieval without reranking, fusion, or pseudo-relevance feedback (Lin et al., 2022). Including these techniques could enhance performance and reveal additional insights into CLIR system design. We also did not analyze the computational costs of translation, retrieval, or indexing, as explored in prior work (Rosa et al., 2021; Nair et al., 2023). Such analysis would be valuable for assessing the trade-offs between effectiveness and efficiency in practical deployment. Additionally, we did not fine-tune or train any retrieval models on our dataset. Training on domain-specific data could potentially lead to better performance, both on our dataset and on others.

Acknowledgments

This project was funded by the Social Science and Humanities Research Council of Canada Pan-Canadian Knowledge Access Initiative Grant (Grant 1007-2023-0001), and the Fonds de recherche du Québec-Société et Culture through the Programme d'appui aux Chaires UNESCO (Grant 338828).

We used computational resources from NodoIA San Francisco (Ministry of Science and Technology of the Province of Córdoba, Argentina).

We thank Érudit for their support and access to data.

References

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. 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, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

Fernanda Beigel and Luciano Digiampietri. 2022. The battle of the languages in national publishing. *Tempo Social, revista de sociologia da USP*, 34(3).

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Aitao Chen. 2002. Cross-language retrieval experiments at clef 2002. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 28–48. Springer.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv:2405.20468*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10885–10897, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, page 338–344, New York, NY, USA. Association for Computing Machinery.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuel Faysse, Patrick Fernandes, Nuno M Guerreiro, António Loison, Duarte M Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H Martins, and 1 others. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

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

Petra Galuščáková, Douglas W. Oard, and Suraj Nair. 2022. Cross-language information retrieval. *arXiv preprint arXiv:2111.05988*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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*, WSDM '23, page 1048–1056, New York, NY, USA. Association for Computing Machinery.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

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)*, pages 26–31, Marseille, France. European Language Resources Association.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Saurabh Khanna, Jon Ball, Juan Pablo Alperin, and John Willinsky. 2022. Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, 3(4):912–930.

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 Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Carlos Lassance, Ronak Pradeep, and Jimmy Lin. 2023. Naverloo@ trec deep learning and neuclir 2023: As easy as zero, one, two, three—cascading dual encoders, mono, duo, and listo for ad-hoc retrieval. In *Proceedings of the Thirty-Second Text REtrieval Conference (TREC 2023). Gaithersburg, Maryland.*

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023a. Overview of the trec 2022 neuclir track. *arXiv preprint arXiv:2304.12367*.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *arXiv preprint arXiv:2404.08071*.

Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. Hc4: A new suite of test collections for ad hoc clir. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 351–366, Berlin, Heidelberg. Springer-Verlag.

Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2023b. Neural approaches to multilingual information retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 521–536, Berlin, Heidelberg. Springer-Verlag.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. 2025. Making text embedders few-shot learners. In *The Thirteenth International Conference on Learning Representations*.

Wing Yan Li, Julie Weeds, and David Weir. 2022. MuSeCLIR: A multiple senses and cross-lingual information retrieval dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1128–1135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Frassetto Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and Xinyu Zhang. 2022. Simple yet effective neural ranking and reranking baselines for cross-lingual information retrieval. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022*, volume 500-338 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual

embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1109–1112, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025. Cross-dialect information retrieval: Information access in low-resource and high-variance languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10158–10171, Abu Dhabi, UAE. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. 2023. Synthetic cross-language information retrieval training data. *arXiv preprint arXiv:2305.00331*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 382–396, Berlin, Heidelberg. Springer-Verlag.

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

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Aurélie Névéol, Antonio Jimeno Yepes, L Neves, and Karin Verspoor. 2018. Parallel Corpora for the Biomedical Domain. In *International Conference on Language Resources and Evaluation*, Miyazaki, Japan. ELRA.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Jiang Niu and Yue Jiang. 2024. Does simplification hold true for machine translations? a corpus-based analysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications*, 11(1):1–10.

Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165–185. Text Mining and Information Analysis of Health Documents.

Janne Pölönen. 2020. Helsinki initiative on multilingualism in scholarly communication.

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.

Guilherme Moraes Rosa, Luiz Henrique Bonifacio, Leandro Rodrigues de Souza, Roberto Lotufo, and Rodrigo Nogueira. 2021. A cost-benefit analysis

of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*.

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France. European Language Resources Association.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Shadi Saleh and Pavel Pecina. 2020. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6849–6860, Online. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. Plaid: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1747–1756, New York, NY, USA. Association for Computing Machinery.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.

Aditya Siddhant, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the International Conference on Machine Learning 2020*, pages 4411–4421.

Felipe Soares and Martin Krallinger. 2019. Bvs corpus: A multilingual parallel corpus of biomedical scientific texts. *arXiv preprint arXiv:1905.01712*.

Felipe Soares, Gabrielli Harumi Yamashita, and Michel Jose Anzanello. 2018. A parallel corpus of theses and dissertations abstracts. In *Computational Processing of the Portuguese Language*, pages 345–352, Cham. Springer International Publishing.

Ian Soboroff. 2023. The better cross-language datasets. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3047–3053, New York, NY, USA. Association for Computing Machinery.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

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

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58(2):713–755.

EM Voorhees. 2005. Trec: Experiment and evaluation in information retrieval.

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

Cuijun Wu, Fei Xia, Louise Deleger, and Imre Solti. 2011. Statistical machine translation for biomedical text: are we there yet? In *AMIA Annual Symposium Proceedings*, volume 2011, page 1290. American Medical Informatics Association.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Bowen Xu, Zhenchang Xing, Xin Xia, David Lo, Qingye Wang, and Shanping Li. 2016. Domain-specific cross-language relevant question retrieval. In *Proceedings of the 13th International Conference on Mining Software Repositories*, pages 413–424.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024a. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2368–2373, New York, NY, USA. Association for Computing Machinery.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024b. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, pages 50–65, Berlin, Heidelberg. Springer-Verlag.

Eugene Yang, Suraj Nair, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Kevin Duh. 2024c. Efficiency-effectiveness tradeoff of probabilistic structured queries for cross-language information retrieval. *arXiv preprint arXiv:2404.18797*.

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for cross-language information retrieval in six less-resourced languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 7–13, Marseille, France. European Language Resources Association.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 99–109, Abu Dhabi, UAE. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward best practices for training multilingual dense retrieval models. *ACM Transactions on Information Systems*, 42(2):1–33.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Jiang Zhu and Haifeng Wang. 2006. The effect of translation quality in MT-based cross-language information retrieval. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 593–600, Sydney, Australia. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore. Association for Computational Linguistics.

A Retrievers

Table 2 provides an overview of the retrievers evaluated in our study.

Inference with neural models was run using 16-bit floating point (fp16) inference on two NVIDIA A30 GPUs, each with 24GB of memory.

BGE-EN-ICL, BGE-m-gemma2, and NV-Embed-v2 require appending task-specific instructions before encoding the queries, which we did following the authors' templates. BGE-EN-ICL (Li et al., 2025) was used in its zero-shot mode, i.e., without in-context examples appended to the queries.

We also experimented with BGE-M3-dense[15] (Chen et al., 2024), which we excluded from the body of the paper because it did not show improved performance or valuable insights.

We implemented BM25 using Pyserini with default parameters and language-specific analyzers (Lin et al., 2021). For PSQ, we used the `fast_psq` implementation by Yang et al. (2024c)[16] with default parameters. We used the English-French matrix trained on 17.6M parallel sentences provided by Yang et al. (2024c).

B Translation

For LLM-based translation, we used a zero-shot prompt inspired by established best practices for

[15] https://hf.co/BAAI/bge-m3

[16] https://github.com/hltcoe/PSQ

Retriever	Type	Pre-train. Lang.	Fine-tuning Lang.	#Params.	Emb. Dim.	Max. Len.
mE5	Dense (single-vector)	Multilingual	Mostly English	560M	1024	512
mGTE-dense			Mostly English, Chinese	305M	768	8192
BGE-m-gemma2			Mostly English, Chinese	9.2B	3584	8192
Solon			French	560M	1024	512
Lajavaness			French-English	560M	1024	512
Croissant		French-English	French-English	1.3B	2048	2048
NV-Embed-v2		Unknown	Mostly English	7.8B	4096	32768
BGE-EN-ICL			Mostly English	7.1B	4096	512
PLAID-X	Dense (multi-vector)	Multilingual	English, French, German, Spanish	560M	128 per token	512
mGTE-sparse	Sparse (Learned)	Multilingual	Mostly English, Chinese	305M	250,000*	8192
BGE-M3-sparse			Mostly English, Chinese	568M	250,000*	8192
SPLADE++		English	English	110M	30,522*	512
BM25	Sparse (Lexical)	–	–	–	49,144*	–
PSQ		–	–	–	715,837*	–

Table 2: Retrievers used in the study. #Params.: Number of parameters. Emb. Dim.: Document embedding dimension. Max. Len.: Maximum number of input tokens allowed by the model. The values in the pretraining and fine-tuning language columns mentioned are approximations; in many cases, intermediate steps are involved, such as initializing from a pretrained model, followed by training with weak supervision and supervised fine-tuning. However, in all cases, fine-tuning data includes some degree of French data. The specific checkpoints used are given in footnotes in section 4.1.
*The embedding dimension of sparse methods is the underlying vocabulary size.

instruction-tuned LLMs[17]. The complete prompt is provided in Table 3. We used sampling with 0.1 temperature and 1.0 top-p.

```
You are a highly skilled translator from French
to English.
Your task is to accurately translate the French
text I provide into English.
You will be provided with a text, and you will
output a JSON object containing the following
information:
{
   translation: string // the translated text
}
Preserve the meaning, tone, and nuance of the
original text.
Please maintain proper grammar, spelling, and
punctuation in the translated version.
```

Table 3: Prompt used for document translation with LLMs. We used a slight variation of this prompt for query translation.

C Statistical signficance

Tables 4 and 5 show the 95% bootstrap confidence intervals for MAP and Recall@100, respectively, for each retrieval method and translation method. For better readability, instead of showing the lower and upper bounds of the confidence intervals, we choose to show which systems are non-significantly different from each other, i.e., the intervals overlap. For example, the MAP interval of PLAID-X+DT with GPT4 overlaps with the interval of NV-Embed-v2+DT with GPT4, but not with the interval of NV-Embed-v2 with no translation (Table 4).

[17] https://docs.anthropic.com/en/prompt-library/polyglot-superpowers, https://platform.openai.com/docs/examples/default-translation.

Retriever	Translation		MAP
BM25	Gold	1	0.611 [2]
SPLADE++	Gold	2	0.609 [1]
PLAID-X	Gold	3	0.605
NV-Embed-v2	Gold	4	0.600
PLAID-X	Docs. (G4)	5	0.586 [6]
NV-Embed-v2	Docs. (G4)	6	0.586 [5]
NV-Embed-v2	None	7	0.580 [8]
PLAID-X	None	8	0.578 [7,9]
NV-Embed-v2	Query (G4)	9	0.575 [8]
SPLADE++	Docs. (G4)	10	0.572 [11,12,13]
PLAID-X	Docs. (L3)	11	0.572 [10,12,13,14]
BGE-m-gemma2	None	12	0.571 [10,11,13,14]
BGE-m-gemma2	Gold	13	0.571 [10,11,12,14]
NV-Embed-v2	Docs. (L3)	14	0.569 [11,12,13]
BGE-m-gemma2	Docs. (G4)	15	0.560
Solon	Gold	16	0.555
BM25	Docs. (G4)	17	0.549 [18,19,20]
SPLADE++	Docs. (L3)	18	0.548 [17,19,20]
PLAID-X	Query (G4)	19	0.548 [17,18,20]
BGE-m-gemma2	Docs. (L3)	20	0.548 [17,18,19]
BGE-m-gemma2	Query (G4)	21	0.543 [22]
NV-Embed-v2	Docs. (Op)	22	0.541 [21,23]
PLAID-X	Docs. (Op)	23	0.539 [22]
Solon	Docs. (G4)	24	0.536 [25]
BGE-EN-ICL	Gold	25	0.535 [24,26]
BGE-m-gemma2	Docs. (Op)	26	0.533 [25,27]
SPLADE++	Docs. (Op)	27	0.530 [26]
mE5	Gold	28	0.526
Solon	Docs. (L3)	29	0.520
Solon	Query (G4)	30	0.516 [31]
BM25	Docs. (L3)	31	0.513 [30,32]
BGE-M3-sparse	Gold	32	0.511 [31]
BGE-EN-ICL	None	33	0.507 [34]
Solon	None	34	0.507 [33]
Solon	Docs. (Op)	35	0.502 [36]
BGE-EN-ICL	Docs. (G4)	36	0.501 [35]
mGTE-dense	Gold	37	0.496
mE5	Docs. (G4)	38	0.490 [39,40]
BM25	Docs. (Op)	39	0.488 [38,40,41,42]
mGTE-sparse	Gold	40	0.487 [38,39,41,42]
BGE-EN-ICL	Docs. (L3)	41	0.486 [39,40,42]
Lajavaness	Gold	42	0.486 [39,40,41]
mE5	Docs. (L3)	43	0.480
Lajavaness	None	44	0.472 [45]
Lajavaness	Docs. (G4)	45	0.470 [44,46]
mGTE-dense	Docs. (G4)	46	0.468 [45]
BGE-M3-sparse	Docs. (G4)	47	0.460 [48,49,50]
mGTE-dense	Docs. (L3)	48	0.459 [47,49,50]
BGE-M3-sparse	Query (G4)	49	0.458 [47,48,50]
Lajavaness	Docs. (L3)	50	0.457 [47,48,49]
Lajavaness	Query (G4)	51	0.454
mGTE-dense	Docs. (Op)	52	0.452 [53]
mGTE-dense	None	53	0.450 [52,54]
mE5	Docs. (Op)	54	0.448 [53]
mGTE-dense	Query (G4)	55	0.445
BGE-EN-ICL	Query (G4)	56	0.441
BGE-M3-sparse	Docs. (L3)	57	0.434 [58,59,60]
mGTE-sparse	Query (G4)	58	0.434 [57,59,60]
mE5	None	59	0.434 [57,58,60]
Lajavaness	Docs. (Op)	60	0.431 [57,58,59]
mGTE-sparse	Docs. (G4)	61	0.428 [62]
SPLADE++	Query (G4)	62	0.426 [61]
BGE-M3-sparse	Docs. (Op)	63	0.413 [64,65]
mE5	Query (G4)	64	0.412 [63,65]
BGE-EN-ICL	Docs. (Op)	65	0.411 [63,64]
mGTE-sparse	Docs. (L3)	66	0.405
mGTE-sparse	Docs. (Op)	67	0.401
BM25	Query (G4)	68	0.390
Croissant	Gold	69	0.376
Croissant	Query (G4)	70	0.365
Croissant	None	71	0.358 [72]
Croissant	Docs. (G4)	72	0.357 [71]
Croissant	Docs. (L3)	73	0.345
Croissant	Docs. (Op)	74	0.325
SPLADE++	None	75	0.284
BM25	None	76	0.181
BGE-M3-sparse	None	77	0.177
mGTE-sparse	None	78	0.169
PSQ	None	79	0.123

Table 4: 95% bootstrap confidence intervals for MAP, using 1000 resamples. Numbers in subscripts indicate the 95% interval of the system of the row overlaps with the interval of the systems in the subscripts.
G4: GPT-4o-mini. L3: Llama-3.2. Op: OpusMT.

Retriever	Translation		Recall@100
BGE-m-gemma2	Gold	1	0.917
BGE-m-gemma2	Docs. (G4)	2	0.910
BGE-m-gemma2	Docs. (L3)	3	0.908
BGE-m-gemma2	None	4	0.903
NV-Embed-v2	None	5	0.895 [6,7,8]
BGE-m-gemma2	Query (G4)	6	0.895 [5,7,8]
BGE-m-gemma2	Docs. (Op)	7	0.894 [5,6,8]
NV-Embed-v2	Gold	8	0.894 [5,6,7,9]
NV-Embed-v2	Docs. (G4)	9	0.892 [8]
NV-Embed-v2	Query (G4)	10	0.889
NV-Embed-v2	Docs. (L3)	11	0.887
PLAID-X	Gold	12	0.879
SPLADE++	Gold	13	0.875 [14]
PLAID-X	Docs. (G4)	14	0.874 [13]
PLAID-X	None	15	0.870 [16,17]
Solon	Gold	16	0.870 [15,17]
PLAID-X	Docs. (L3)	17	0.869 [15,16]
Solon	Docs. (G4)	18	0.866 [19]
NV-Embed-v2	Docs. (Op)	19	0.866 [18,20]
SPLADE++	Docs. (G4)	20	0.864 [19]
BGE-EN-ICL	Gold	21	0.861 [22,23]
BM25	Gold	22	0.861 [21,23]
Solon	Docs. (L3)	23	0.860 [21,22,24]
Solon	Query (G4)	24	0.858 [23,25,26]
BGE-EN-ICL	None	25	0.857 [24,26]
Solon	None	26	0.856 [24,25,27]
PLAID-X	Query (G4)	27	0.854 [26,28]
SPLADE++	Docs. (L3)	28	0.853 [27]
mGTE-dense	Gold	29	0.849 [30,31]
Lajavaness	Gold	30	0.849 [29,31]
Lajavaness	None	31	0.848 [29,30]
PLAID-X	Docs. (Op)	32	0.845 [33]
Solon	Docs. (Op)	33	0.845 [32,34]
Lajavaness	Docs. (G4)	34	0.843 [33]
mE5	Gold	35	0.840
Lajavaness	Query (G4)	36	0.838 [37,38,39,40]
mGTE-dense	Docs. (G4)	37	0.837 [36,38,39,40]
BGE-EN-ICL	Docs. (G4)	38	0.837 [36,37,39,40]
Lajavaness	Docs. (L3)	39	0.836 [36,37,38,40,41]
SPLADE++	Docs. (Op)	40	0.836 [36,37,38,39,41]
mGTE-dense	Docs. (L3)	41	0.834 [39,40,42]
BM25	Docs. (G4)	42	0.832 [41,43]
BGE-EN-ICL	Docs. (L3)	43	0.831 [42]
mE5	Docs. (G4)	44	0.823
mGTE-dense	None	45	0.820 [46]
mGTE-dense	Docs. (Op)	46	0.820 [45]
mE5	Docs. (L3)	47	0.817 [48,49]
Lajavaness	Docs. (Op)	48	0.817 [47,49]
BM25	Docs. (L3)	49	0.815 [47,48,50]
mGTE-dense	Query (G4)	50	0.813 [49]
BGE-EN-ICL	Query (G4)	51	0.810
BGE-M3-sparse	Gold	52	0.807 [53]
mGTE-sparse	Gold	53	0.805 [52]
Croissant	Query (G4)	54	0.794 [55,56]
Croissant	Gold	55	0.794 [54,56]
Croissant	None	56	0.793 [54,55]
mE5	Docs. (Op)	57	0.790 [58]
BM25	Docs. (Op)	58	0.789 [57]
mE5	None	59	0.784
Croissant	Docs. (G4)	60	0.781 [61]
BGE-M3-sparse	Query (G4)	61	0.781 [60]
BGE-M3-sparse	Docs. (G4)	62	0.778
Croissant	Docs. (L3)	63	0.773 [64]
mGTE-sparse	Docs. (G4)	64	0.771 [63]
mGTE-sparse	Query (G4)	65	0.763 [66,67]
BGE-M3-sparse	Docs. (L3)	66	0.763 [65,67,68]
mGTE-sparse	Docs. (L3)	67	0.760 [65,66,68,69]
BGE-EN-ICL	Docs. (Op)	68	0.760 [66,67,69]
mE5	Query (G4)	69	0.760 [67,68,70]
PSQ	None	70	0.757 [69]
SPLADE++	Query (G4)	71	0.753
Croissant	Docs. (Op)	72	0.748
BGE-M3-sparse	Docs. (Op)	73	0.738 [74]
mGTE-sparse	Docs. (Op)	74	0.737 [73]
BM25	Query (G4)	75	0.706
SPLADE++	None	76	0.604
BGE-M3-sparse	None	77	0.449
mGTE-sparse	None	78	0.443
BM25	None	79	0.417

Table 5: 95% bootstrap confidence intervals for Recall@100, using 1000 resamples. Numbers in subscripts indicate the 95% interval of the system of the row overlaps with the interval of the systems in the subscripts
G4: GPT-4o-mini. L3: Llama-3.2. Op: OpusMT.