
BRIDGING LANGUAGE GAPS: ADVANCES IN CROSS-LINGUAL INFORMATION RETRIEVAL WITH MULTILINGUAL LLMs

Roksana Goworek*
The Alan Turing Institute
Queen Mary University of London

Olivia Macmillan-Scott*
The Alan Turing Institute
University College London

Eda B. Özyiğit
The Alan Turing Institute

{rgoworek, omacmillan-scott, eozyigit}@turing.ac.uk

ABSTRACT

Cross-lingual information retrieval (CLIR) addresses the challenge of retrieving relevant documents written in languages different from that of the original query. Research in this area has typically framed the task as monolingual retrieval augmented by translation, treating retrieval methods and cross-lingual capabilities in isolation. Both monolingual and cross-lingual retrieval usually follow a pipeline of query expansion, ranking, re-ranking and, increasingly, question answering. Recent advances, however, have shifted from translation-based methods toward embedding-based approaches and leverage multilingual large language models (LLMs), for which aligning representations across languages remains a central challenge. The emergence of cross-lingual embeddings and multilingual LLMs has introduced a new paradigm, offering improved retrieval performance and enabling answer generation. This survey provides a comprehensive overview of developments from early translation-based methods to state-of-the-art embedding-driven and generative techniques. It presents a structured account of core CLIR components, evaluation practices, and available resources. Persistent challenges such as data imbalance and linguistic variation are identified, while promising directions are suggested for advancing equitable and effective cross-lingual information retrieval. By situating CLIR within the broader landscape of information retrieval and multilingual language processing, this work not only reviews current capabilities but also outlines future directions for building retrieval systems that are robust, inclusive, and adaptable.

Keywords Cross-lingual information retrieval · Multilingual large language models · Cross-lingual embeddings · Retrieval evaluation methods

1 Introduction

Given a query and a set of documents, information retrieval (IR) [1–3] is the task of identifying documents that are relevant to the query. Cross-lingual information retrieval [4–7] extends this task by enabling queries expressed in one language to retrieve documents written in one or more different languages. Unlike traditional monolingual IR, which assumes a shared language between query and documents, CLIR faces the challenge of bridging language boundaries. This is typically addressed by combining techniques from both information retrieval and multilingual natural language processing (NLP).

The emergence of the Internet and search engines in the 1990s revealed striking disparities in linguistic accessibility. At that time, English accounted for almost 80% of all web content, although it was the native language of only a small share of users [8, 9]. This imbalance underscored the need for research into cross-lingual information access. For example, high-resource languages such as English, Spanish, and Chinese developed strong web presence and benefited from early NLP support, while many others, particularly low-resource languages like Swahili or Burmese, lagged behind due to limited digital content and inadequate computational tools. These disparities persist today, with most modern web content and NLP systems still skewed towards a few dominant languages [10]. The rise of LLMs has only amplified this imbalance, as most high-performance models remain disproportionately trained and optimised for

*Equal contribution.

English; English accounts for around 90% of the training data in most popular models [11]. CLIR offers a compelling response by enabling users to access information written in other languages, thereby helping to democratise knowledge across linguistic boundaries.

Traditional monolingual information retrieval systems are generally organised into a multi-stage pipeline: (i) query expansion, which broadens the query using synonyms, spelling corrections, or related terms to improve recall; (ii) ranking, which performs an efficient first-pass retrieval to select a candidate set of relevant documents; (iii) re-ranking, which applies more computationally intensive models to refine the order of the top documents; and optionally (iv) generation, where an answer or summary is synthesised, often by large language models. Cross-lingual retrieval follows the same pipeline while introducing additional complexity, for example by translating the query and/or the documents, aligning multilingual embeddings for semantic similarity, or leveraging generative multilingual models to bypass translation altogether.

Recent advances in neural language modelling, particularly in cross-lingual embedding and multilingual pre-training, have enabled more powerful and flexible architectures for CLIR. These systems increasingly use multilingual sentence encoders [12], dense retrieval [13–15], and large-scale generative language models [16] to compare aligned representations across languages. With improved multilingual corpora and evaluation datasets, systems are becoming more effective, scalable, and easier to benchmark. Yet building reliable cross-lingual retrieval remains difficult. Considerable research still treats CLIR as monolingual retrieval plus translation, which oversimplifies the problem and overlooks multilingual challenges. Lexical, syntactic, and semantic differences hinder alignment, and many language pairs lack parallel corpora, especially low-resource languages. Domain mismatch further reduces generalisability. Constructing high-quality, gold-standard judgments is resource-intensive, while annotation transfer introduces noise and inconsistency. Translation and alignment may cause semantic drift, altering meaning and reducing accuracy. To overcome these issues, next-generation systems must go beyond translation-based pipelines and address linguistic, resource, and evaluation mismatches directly. Although there are surveys of information retrieval [17–19] and multilingual NLP [20–23], few focus on CLIR, and none fully examine embedding-based retrieval. Early work, such as that by Nie [24], focused on translation-based methods, but more recent developments [25] highlight embedding-based retrieval and generative models, which remain in the early stages of adoption.

This survey provides a comprehensive overview of recent advances across the full pipeline (from query reformulation to re-ranking and answer generation), with a focus on multilingual embedding alignment, contrastive learning, multilingual pre-training strategies, and the integration of generative language models for retrieval and response generation (see Figure 1). It presents a unified perspective of current methods, available resources, and the open challenges in cross-lingual system design. Its key contributions are:

- **CLIR techniques.** An analysis of cutting-edge CLIR methods, including embedding alignment, multilingual pre-training, multilingual LLMs, and retrieval architectures, highlighting their strengths and trade-offs.
- **Multilingual advances.** A review of recent developments in multilingual NLP and their relevance to CLIR showing how progress in cross-lingual models supports multilingual information retrieval.
- **Datasets and evaluation.** A survey of datasets, evaluation protocols, and performance metrics, along with a discussion of current limitations and opportunities for improvement.
- **Core challenges.** An analysis of major challenges in CLIR, including linguistic divergence, data scarcity, domain and language generalisation, and fairness, together with their implications for real-world deployment. This identifies obstacles to practical adoption and highlights areas needing further research.

This work is organised around two central dimensions: (i) how systems implement core components of the retrieval process, and (ii) how they address cross-linguality. Section 2 introduces system architectures and integration of cross-lingual representations into scalable retrieval pipelines. Section 3 examines strategies for cross-linguality, including translation-based methods, multilingual LLMs, embeddings, and alignment techniques. Section 4 reviews evaluation practices, focusing on benchmark datasets, performance metrics, and the need for fair multilingual assessment. Section 5 explores real-world applications such as multilingual search, cross-lingual question answering (QA), and domain-specific information access. Section 6 discusses key challenges such as linguistic divergence, resource scarcity, evaluation difficulties, model limitations, and highlights future research directions. Finally, section 7 concludes with a synthesis of insights and a discussion of the broader impact of cross-lingual retrieval.

2 CLIR Architecture

While the primary challenge in CLIR lies in bridging language gaps, it is equally important to consider how the core architecture of IR systems can be utilised and extended. This section focuses on the foundational stages of the retrieval

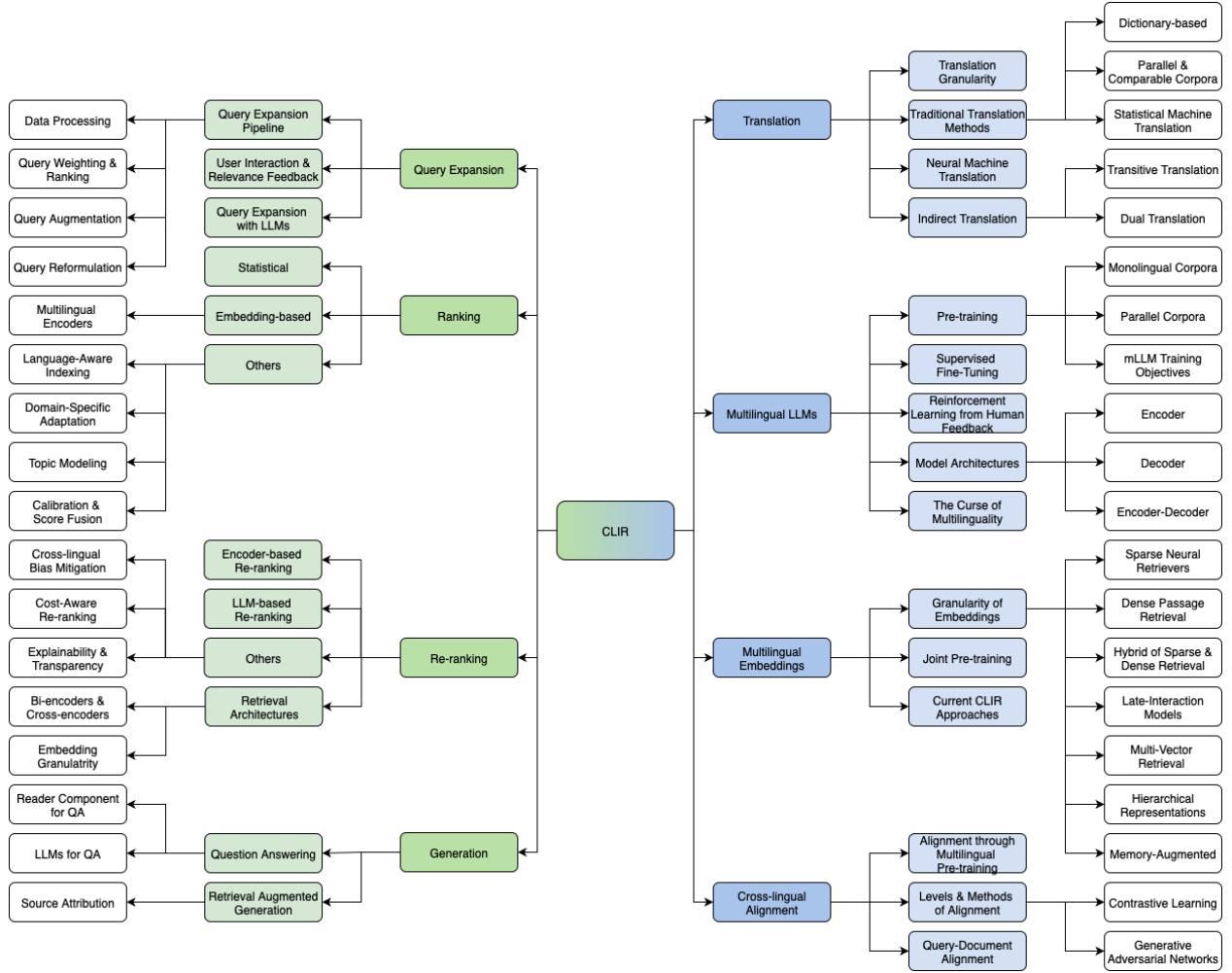


Figure 1: Overview of the survey covering cross-lingual and information retrieval approaches.

pipeline: query expansion, ranking, re-ranking, and, optionally, question answering. It highlights recent advances in monolingual IR that have the potential to strengthen CLIR performance rather than addressing how systems handle cross-linguality itself. Figure 2 illustrates how the query "origin of dumplings" may be processed through the CLIR architecture: beginning with query expansion, moving to initial ranking and re-ranking, followed by question answering, and culminating in a final response generated by a LLM.

2.1 Query Expansion

The effectiveness of information retrieval relies on the assumption that the user’s initial query accurately reflects their information requirements. However, this assumption is often too strong. In web search, several studies have found that

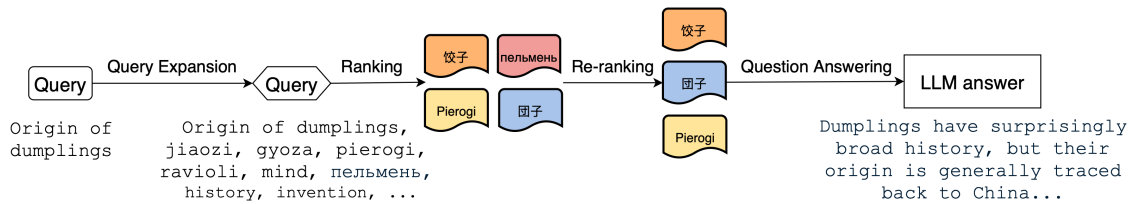


Figure 2: CLIR architecture example. The query "origin of dumplings" passes through expansion, translation and/or embedding before relevant documents are retrieved and re-ranked.

the average length of queries is less than 2.5 words [26, 27]. Such short queries provide insufficient context and leave substantial room for ambiguity. Users may also misspell terms, resulting in lower retrieval performance. In the case of cross-lingual retrieval, additional difficulties arise, as there may be multiple possible translations or transliterations of a given word.

Another issue is the *vocabulary problem* [28]: the terms used in a query may not match those used in the documents themselves or terms used to index the documents. As a result, lexical-based retrieval approaches can fail to identify relevant documents. Information retrieval faces the added difficulty of synonyms and word inflections, which can lower recall [29]. Similarly, the concept of polysemy, where a single word may contain multiple meanings, can further reduce retrieval performance.

Query expansion provides an effective method for addressing these issues, particularly in cases of ambiguity or lower-quality queries. The user’s initial query is expanded using relevant terms or synonyms, in so doing improving the retrieval performance and more closely aligning with the user’s information needs. For example, Figure 2 demonstrates how the query "origin of dumplings" may be expanded with terms such as *gyoza* or *history*. Query expansion is especially relevant for what Broder [30] denotes as informational queries, which are broad in scope and often correspond to many related documents. In contrast, navigational queries typically seek a specific result, and transactional queries involve the intent to perform a particular activity.

Although query expansion is the more common strategy for overcoming the limitations of short or underspecified queries, recent research has also explored expanding the documents themselves. One example is Nogueira et al.’s [31] Doc2query approach, in which a set of possible queries is predicted for each document and appended as pseudo-queries. The expanded document is then indexed and used in retrieval with traditional methods such as BM25 [32]. In practice, some document expansion techniques have proven more effective than query expansion, particularly in sparse retrieval settings [33].

2.1.1 Query Expansion Pipeline

The query expansion pipeline can be divided into four components: preprocessing of the data source, term weighting and ranking, expansion term selection, and query reformulation [29, 34]. The main challenges are determining which terms to use, how to weight them, and how to integrate them into the query [24].

Data Processing. In the first stage, potential expansion terms are extracted from the chosen data source. These sources may include thesauri and ontologies such as WordNet [35], Wikipedia datasets, search and query logs, and word embeddings. Relevance feedback [36, 37] is another widely used strategy: feedback may be global, drawing on the entire collection, or local, relying on the set of documents initially retrieved. Xu and Croft [38] note that local feedback methods are often more effective, since they exploit context-specific evidence rather than general collection statistics. LLMs have also been used to generate pseudo-documents, thereby enriching the document side of retrieval.

Query Weighting and Ranking. Once the relevant data sources have been processed to collect potential expansion terms, these are weighted and ranked to determine the most appropriate ones. Relevant terms are selected, either through lexical approaches to identify potential synonyms, hypernyms, or words identified through statistical or semantic similarity. Approaches differ in how they model the relationship between original query terms and candidate terms [29]. One-to-one associations link each query term directly to an expansion term, often using stemming [39] or thesauri. By contrast, one-to-many associations identify terms related to multiple query terms simultaneously. For example, through co-occurrence analysis [40, 41]. Other methods apply statistical or model-based weighting techniques, in which probabilistic or language models estimate the strength of association between query and candidate terms [42, 43].

Query Augmentation. Having created a ranked or weighted set of possible expansion terms, the next step is to determine which of those will be used to augment the query. Some have argued that having a smaller number of expansion terms is beneficial as it reduces the noise that can be introduced into the query [44], whereas others have claimed that the quality of the selected terms is more important than how many of them are used [45]. Instead of trying to determine how many terms is the optimum number, others have achieved higher performance by employing more informed selection techniques [46, 47]. Nevertheless, the number of expansion terms proposed has varied widely in the literature, from 5-10 terms [48, 49], to a few hundred terms [50–52], to a third of all candidate terms [53].

Query Reformulation. In the final stage, the selected expansion terms are incorporated into the original query to produce an augmented version [44]. The objective is to capture the user’s information need more precisely and to mitigate the vocabulary problem.

2.1.2 User Interaction & Relevance Feedback

User involvement in query expansion varies by approach. Automatic query expansion [29] requires no input, while manual and interactive methods involve users to prevent concept drift [54]. In manual expansion, users directly select terms, often with query log support. Interactive query expansion offers system-suggested terms from which users choose.

Relevance feedback provides another interaction mechanism. Early work by Rocchio [36, 37], based on the SMART retrieval system [55, 56], uses user-identified relevant documents to expand queries through local feedback expansion, which proved more effective than global, resource-based methods.

Pseudo-relevance feedback (PRF), introduced by Croft and Harper [57], employs a similar approach but removes user input by assuming the top-ranked documents are relevant. Though generally less robust than explicit feedback, PRF provides advantages as it is fully automated and so is widely adopted. RM3 [58], an influential PRF algorithm, builds on Lavrenko and Croft’s relevance model [42] by reintroducing original query terms back into the expansion.

2.1.3 Query Expansion with LLMs

The development of neural network-based language models has opened up new possibilities for query expansion [59]. Zhu et al. [18] classify LLM-based approaches into three main categories: prompting, supervised fine-tuning and reinforcement learning. Li et al. [59] further consider alignment techniques such as preference optimisation and distillation. Prompting methods, where models directly generate expanded queries or related documents, include zero-shot, few-shot, and Chain-of-Thought (CoT) techniques. Fine-tuning requires large task-specific datasets, and has so far seen limited use in query expansion [60]. Reinforcement learning allows models to refine expansions using feedback from retrieval systems. LLMs typically generate reformulated queries (e.g. rephrasings or keyword expansions [61]), concept-based queries, or answer-enriched queries that bridge the semantic gap between short queries and long documents [18].

Most research on LLM-based query expansion has focused on prompting. For instance, Clauveau [62] applies zero-shot prompts to generate multiple expansions, concatenated with the original query for retrieval via BM25+ [63]. Similarly, Gao et al. [64] generate pseudo-documents using zero-shot prompting, later embedded for dense retrieval. Query2doc also produces pseudo-documents, which can then be concatenated with the query for both sparse and dense retrieval [33]. Hypothetical Document Embeddings (HyDE) [64] adopts a comparable approach focused on dense retrieval. Few shot-prompting, also known as in-context learning, has also been used in the Query2doc method to generate pseudo-documents which are then concatenated to the original query, used for both sparse and dense retrieval [33]. A comparative study found CoT prompting to be the most effective [65].

Despite these advances, a key challenge lies in mitigating hallucinations (i.e. irrelevant or erroneous expansions that degrade retrieval performance). Abe et al. [66] attribute these issues to knowledge gaps and ambiguous prompts, while other work suggests consistency verification to filter meaningless outputs [67]. LLMs offer another potential approach to address the vocabulary problem and ambiguity in queries through conversation search, where user interaction can be used to clarify their intent. This helps mitigate concept drift, especially when models generate long-term pseudo-documents.

Research on multilingual query expansion remains limited. Recent work often combines translation and expansion, performed sequentially or jointly [68, 69]. Implicit expansion, where vector representations are adjusted instead of adding related terms [59], has shown promise. Query expansion in CLIR can both improve alignment with user intent and reduce translation-induced errors [70]. Sequential pre-expansion translation has been shown to be effective, though hybrid pre- and post-expansion methods often yield the best performance [71, 72].

2.2 Ranking

The ranking stage retrieves an initial list of documents from the collection and scores them based on relevance to the query. This stage prioritises efficiency and recall, providing a coarse-grained ordering that can later be refined during re-ranking. Ranking methods fall into three categories: (i) traditional statistical approaches; (ii) embedding-based neural retrieval; and (iii) hybrid or enhanced strategies that incorporate language awareness and fine-tuning.

2.2.1 Traditional Statistical Ranking

Statistical methods such as Term Frequency-Inverse Document Frequency (TF-IDF) [73] and BM25 [32] form the historical foundation of CLIR. These approaches rely on exact term matching, making them fast, interpretable, and computationally efficient. BM25, in particular, has become a de facto baseline due to its robustness across domains.

However, these models require the query and document to share a language, limiting their direct use in CLIR without translation. Extensions with translation enable their application in cross-lingual contexts, where they remain competitive in low-resource or high-latency settings.

2.2.2 Neural Embedding-based Ranking

With the rise of multilingual pre-trained language models, embedding-based retrieval has become dominant in CLIR. These models map queries and documents into a shared vector space, enabling retrieval via similarity metrics such as cosine or dot product rather than lexical overlap.

A common architecture is the bi-encoder, also known as a dual-encoder, in which queries and documents are independently encoded into fixed-size embeddings. The document embeddings can be pre-computed and indexed, making retrieval highly efficient even over large corpora. These and other ranker and re-ranker architectures are described in Section 2.3 as there is significant overlap in the techniques used. Similar models can be used for both ranking and re-ranking, however, ranking requires fast retrieval over extremely large corpora, making only efficient methods applicable.

The performance of multilingual embeddings can sometimes be improved by incorporating language-specific information. This helps filter irrelevant matches and reduce semantic drift in multilingual corpora. One strategy is to augment document or query embeddings with language identifiers or metadata [74, 75]. By constraining retrieval to the appropriate linguistic space, these methods reduce noise and increase precision.

General-purpose multilingual models often underperform in specialised domains, where vocabulary and semantic relations differ from open-domain text. Domain-specific adaptation addresses these challenges by tailoring retrieval models to the target field. Methods include training with domain-specific positive and negative pairs to improve semantic alignment [76], continued pre-training on in-domain corpora to expand coverage of specialised vocabulary [77], and instruction tuning with domain-relevant tasks to better capture real information needs [78]. These strategies enhance retrieval accuracy in areas such as law, medicine, and technology, where precision and contextual understanding are critical.

Before the widespread adoption of neural embeddings, CLIR systems frequently employed topic modelling to capture latent semantics across languages. Latent Dirichlet Allocation (LDA) and multilingual extensions grouped documents into topics, providing a coarse but effective basis for cross-lingual retrieval. Further developments such as Bilingual LDA (BiLDA) [79] and Polylingual Topic Models (PLTM) [80] aligned topic spaces across languages using dictionaries or parallel corpora, while supervised models like Polylingual Labelled LDA [81] improved alignment with labelled data. Although largely supplanted by neural embeddings, topic modelling remains useful in low-resource environments or as a complementary component in hybrid systems.

In many retrieval scenarios, particularly web search and e-commerce, performance depends on features beyond query-document similarity. Query properties such as length, phrasing, and intent are strong predictors of retrieval success [82]. Structural features, including PageRank, freshness (i.e. a temporal feature that measures how recent a document is relative to the query or current time), and categorical tags, further refine ranking. Behavioural signals such as clicks, ratings, and purchases, serve as implicit relevance feedback, often underperforming purely textual features. Learning-to-rank methods like LambdaMART [83] incorporate these signals effectively, and normalisation strategies [84] help reduce bias, making feature-enhanced ranking especially valuable in multilingual applications.

2.3 Re-ranking

After an initial retrieval by a first-stage ranker, re-ranking refines the ordering of candidate documents using more detailed relevance signals. Since it applies only to a limited subset (e.g. top 100 results), computationally intensive models can be used, often incorporating semantic analysis, external signals, or joint ranking strategies. Re-ranking is particularly important in high-stakes or user-facing CLIR systems.

Models differ in how they formulate the relevance estimation objective. The main paradigms are pointwise, pairwise, and listwise, each balancing complexity, interpretability, and optimisation goals. In pointwise approaches, each document is scored independently via regression or classification [85, 86]. This simple and scalable setup is suited to early-stage applications or when labelled data is scarce, but it may miss fine-grained distinctions. Pairwise models compare document pairs, predicting which is more relevant; methods such as RankNet [87] and LambdaMART [88] remain widely used for the ability to handle noisy or ordinal labels. However, they may not capture global ranking structures as effectively as listwise methods. Listwise approaches optimise over the entire ranked list using loss functions such as ListNet [89] and ListMLE [90], with recent transformer-based extensions further improving alignment with evaluation metrics like Normalised Discounted Cumulative Gain (nDCG) [91] or Mean Average Precision (MAP) [92]. Though

more complex, listwise models better capture interdependencies and are particularly effective for long-context or diversity-sensitive ranking tasks [93].

Encoder-based Re-ranking. Encoder-based methods form the foundation of neural ranking in CLIR, primarily realised as cross-encoders and bi-encoders, explained in detail in Section 2.3. Unlike bi-encoders, which encode queries and documents separately, cross-encoders concatenate them into a single sequence so that the transformer’s attention layers can model fine-grained token-level interactions. For this reason, bi-encoders are often used for initial ranking, whereas cross-encoders are preferred for re-ranking. The latter typically yields higher precision but incurs substantial computational costs, limiting scalability to large candidate sets. Multilingual pre-trained models such as XLM-R [94], mT5 [95], and MiniLM [96] have been used as cross-encoders in CLIR, fine-tuned on datasets like multilingual MS MARCO [97], XOR-Retrieve [98], or MIRACL [99], achieving strong language-agnostic ranking. The encoders are often trained with contrastive objectives such as Multiple Negatives Ranking Loss or Margin Ranking Loss, allowing large-scale retrieval with competitive accuracy.

In knowledge-distillation frameworks such as Translate-Distill [100], a cross-encoder acts as a teacher by producing high-quality relevance scores, while a bi-encoder serves as a student, learning to approximate those scores [101, 102]. This allows the final bi-encoder model to achieve efficient dense retrieval with accuracy close to the more computationally expensive cross-encoder [100].

LLM-based Re-ranking. Language models have recently emerged as powerful re-rankers, leveraging advanced semantic understanding that often surpasses traditional cross-encoders, particularly in zero-shot and few-shot scenarios. By reasoning directly over candidate lists, LLMs can improve retrieval effectiveness without requiring task-specific fine-tuning, though challenges such as inference cost, latency, and prompt sensitivity remain.

- *Prompt-based re-ranking.* Prompting LLMs directly enables document scoring (pointwise), pairwise comparisons, or listwise reordering. Frameworks such as HyDE [64], InPars [103], or RankGPT[104] adopt this approach. Pairwise ranking prompting has been shown to outperform GPT-4 pointwise re-ranking in some cases.
- *Zero-shot & few-shot listwise re-ranking.* LLMs can reorder candidates even without specific fine-tuning, often surpassing cross-encoders in zero-shot and few-shot settings. Ma et al. [105] propose a zero-shot listwise re-ranker that achieves strong nDCG gains, particularly in multilingual datasets. Few-shot prompting further improves performance as shown by PaRaDe [106].
- *Document entailment & instruction-tuned models.* Instruction-tuned LLMs like Flan-T5 [107], Zephyr [108], and ChatGPT [109] demonstrate strong capability in assessing document relevance. For instance, Flan-T5 XXL outperforms baselines when treated as an entailment verifier.
- *Attention-based ranking (in-context re-ranking).* In-context re-ranking (ICR) leverages transformer attention to rank documents without generating full text. Chen et al. [110] show that ICR reduces latency by over 60% compared to generative prompting while maintaining accuracy.
- *Cross-encoder vs LLM re-rankers.* While LLMs like GPT-4 achieve impressive zero-shot performance, cross-encoders remain competitive in domain-matched settings [111]. This suggests that LLMs are promising but not yet universally dominant.

Additional Approaches. While cross-encoders, bi-encoders, and other neural re-ranking architectures remain central to CLIR, several complementary strategies have been explored. These approaches address challenges such as efficiency, bias, interpretability, and integration with downstream tasks, providing practical extensions beyond the core paradigms.

- *QA-oriented re-ranking.* In CLIR pipelines aimed at question answering, retrieval can be optimised for answerability rather than raw relevance. The MIA shared task [112] applied a zero-shot multilingual question-generation model to top-k passages, scoring by the probability of regenerating the query. This eliminated the need for annotated data and outperformed BM25 by 6–18% in top-20 accuracy.
- *Pre-trained and joint QA models.* Pre-trained extractive QA models act as implicit cross-language re-rankers by extracting answer spans. They assess the likelihood of a passage supporting the correct answer by extracting start/end positions. This approach has been validated on multilingual QA datasets like MLQA [113] and XOR-TyDi QA [98], where re-ranking based on QA model confidence yields higher answer recall and precision [98].
- *Feature-enhanced re-ranking.* Industrial systems often incorporate structural and behavioural features alongside textual similarity. Features such as freshness, domain authority, click-through rates, and user engagement

improve ranking utility. Feature-aware models (e.g. LambdaMART [114]) yield notable gains, especially in low-resource or underrepresented languages where semantic alignment is noisy.

- *Score calibration and fusion.* High quality re-ranking often requires integrating outputs from multiple retrieval components (BM25, dense encoders, cross-encoders). Reciprocal Rank Fusion [115] avoids score-scale mismatches and is used in Zilliz [116] and Azure AI [117]. Beyond fusion, calibration techniques (e.g. linear interpolation, normalisation) prevent biases from dominant languages or models.
- *Mitigating cross-lingual bias.* Language-specific scoring biases arise when source-language documents receive inflated relevance scores Zhang et al. [84]. Mitigation strategies include (i) score normalisation [118]; (ii) balanced training with models such as LambdaMART [88, 114]; and (iii) adaptive thresholds [119]. Positional bias is further addressed by the Cascade Model/UBM click models and re-ranking with RandPair or FairPair [120–123]. For LLM-based re-ranking, prompt-shuffling or permutation improves output stability [124, 125]. To avoid repeated content or excessive focus on a single topic, methods like Maximal Marginal Relevance selectively promote new information while maintaining relevance. Early work [126] demonstrated improved information coverage using this method in both retrieval and summarisation contexts. Lin et al. [127] show that explicit novelty scoring yields better user satisfaction in multilingual CLIR systems.
- *Cost-aware re-ranking.* To reduce computational cost, cascading approaches filter candidates before applying expensive re-rankers. Bi-encoder cascades cut compute by up to sixfold with minimal performance loss [128–130]. Adaptive candidate truncation for language model-based re-rankers [131] further optimises the trade-off between efficiency and retrieval quality.
- *Dynamic and adaptive re-ranking.* Reinforcement learning and feedback-driven re-rankers dynamically adapt rankings to user interactions. RLIRank [132] uses reinforcement learning with Long Short-Term Memory (LSTM) [133] based click modelling, outperforming static rankers. Studies [134, 135] demonstrate that reinforcement feedback enhances performance on dynamic tasks such as Text REtrieval Conference (TREC) Dynamic Track [136], even with limited supervision.
- *Explainability and transparency.* In high-stakes domains (e.g. legal or medical), interpretable ranking is essential. Approaches include extractive explanations (Select-And-Rank [137]), attention-based rationales [138, 139], and post-hoc explanation platforms. Frameworks such as Stable and Explainable Attention (SEAT) [140] align attention and predictions, while benchmarks [141, 142] stress balancing performance with interpretability, accountability, and auditability in CLIR deployment.

Retrieval Architectures. There are many different architectures which can be used for the ranking and re-ranking stages of the CLIR pipeline. Bi-encoders, the most common architecture for ranking, encode queries and documents into dense vectors using a shared transformer backbone, estimating relevance via similarity functions such as dot product or cosine similarity. Document embeddings are precomputed offline, while query embeddings are generated online and matched using the similarity function. They do not allow token-level interaction but are more efficient since embeddings can be precomputed. This design allows efficient and scalable retrieval, but struggles with fine-grained text interactions. Cross-encoders, by contrast, concatenate query and document inputs and process them jointly within a transformer model with full cross-attention, capturing detailed query-document interactions and improving ranking accuracy. However, each query-document pair must be processed individually at inference, making this approach computationally expensive and therefore typically reserved for re-ranking. In practice, systems often combine both approaches: bi-encoders for first-stage retrieval over large collections, and cross-encoders for re-ranking a smaller candidate set.

More recent models, such as late interaction methods and sparse neural retrievers, have been designed specifically for information retrieval and often outperform traditional approaches. To balance efficiency and accuracy, late-interaction models (e.g. ColBERT [143] and ColBERTv2 [144]) separately encode queries and documents, then compute token-level similarity via MaxSim [145] (i.e. compute the sum of maximum similarities between each query token and all document tokens), enabling fine-grained matching at reduced cost. Sparse neural retrievers, meanwhile, exploit sparse inverted-index efficiency while incorporating semantic richness. Approaches include term-weighting methods such as DeepCT [146] and expansion models such as SPLADE [147, 148] and SpaDE [149], which use transformers to predict term importance or generate additional relevant terms. These maintain compatibility with inverted indices while introducing semantic depth.

Each architecture offers distinct trade-offs. Bi-encoders enable scalability and high efficiency but may miss nuanced semantics. Cross-encoders capture complex interactions but are computationally prohibitive for large-scale retrieval. Late-interaction models offer a middle ground, retaining token-level richness with manageable overhead. Sparse retrievers combine the efficiency of inverted indices with neural modelling, making them well suited for large-scale tasks. The effectiveness of these models also depends on embedding granularity, whether at the word, sentence, passage,

Approach	Granularity	Representation	Retrieval Level	Context Modelling
Sparse Retrieval	Term-level	High-dimensional sparse vectors (e.g. BM25, SPLADE)	Term matching via inverted index	Encoded term importance; no dense semantic context [147, 151]
Dense Passage Retrieval	Passage-level (100–300 words)	One embedding per passage	Passage retrieval	Independent encoding of each passage without cross-passage context
Hybrid Sparse + Dense Retrieval	Term + Document (via dense)	Concatenated or parallel sparse and dense vectors	Retrieval via fusion or unified indexing	Sparse term precision + dense semantic bridging between languages [153, 154]
Late Interaction Models (e.g. ColBERT)	Token-level embeddings	Contextualised token embeddings	Token-level MaxSim retrieval	Combines local token interactions with global context via transformer encoders
Multi-Vector Retrieval Models (e.g. ME-BERT, COIL)	Token- or span-level	Multiple dense vectors per document (token/span)	Token or span matching	Combines fine-grained token matching with optional global encoding
Hierarchical Representations	Multi-level: Word → Sentence → Paragraph	Hierarchical combination of embeddings	Document or segment retrieval	Captures local structure and aggregates into global document context
Memory-Augmented Models	Embeddings for document parts (e.g. passages, sentences)	Memory slots for different parts	Retrieval via memory attention	Query dynamically attends to relevant parts of the document

Table 1: Comparison of document representation and retrieval approaches at different embedding granularities

or document level. Short queries are often represented well by a single vector, but longer texts demand finer-grained embeddings. Traditional sparse methods like BM25 [150] rely on term-frequency statistics (e.g. TF-IDF [73]), surfacing documents through keyword overlap, but embedding-based models risk collapsing diverse topical content into a single dense vector, overlooking relevant information. To mitigate this, fine-grained approaches have been developed.

Sparse retrieval models such as BM25, SPLADE, SPLADE-X represent queries and documents as high-dimensional sparse vectors, enabling inverted-index lookup while incorporating learned expressions (e.g. cross-lingual mappings [147, 151, 152]) to improve CLIR effectiveness. Dense Passage Retrieval (DPR) embeds text segments (100–300 words) independently and has shown substantial improvements over BM25, though performance depends on segmentation. Hybrid models combine sparse and dense signals [153, 154], either through parallel retrieval with later rank fusion (e.g. Reciprocal Rank Fusion [115, 154] or concatenation, which often outperforms either approach alone).

Late-interaction models like ColBERT and ColBERTv2 [143, 144] retain token-level embeddings and compute fine-grained similarity at the cost of higher storage requirements. Multi-vector retrieval methods (e.g. ME-BERT [155], COIL [156]) similarly encode documents into multiple vectors for semantic matching, offering strong performance on nuanced queries but demanding greater storage and computation.

Further refinements include hierarchical representations, such as Dense Hierarchical Retrieval [157], which retrieves at the document-level and refines at the passage level, preserving both global and local context. Memory-augmented architectures (e.g. EMAT [158], MoMA [159]) store distinct document segments in explicit memory slots, enabling selective attention during retrieval and dynamic external memory access. These enhance performance but introduce added complexity and computational overhead.

In summary, retrieval architectures vary across dual, cross, late-interaction, sparse, hybrid, hierarchical, and memory-augmented models, each offering trade-offs among scalability, semantic depth, precision, and efficiency. The choice of embedding granularity and retrieval mechanism should ultimately align with the task’s demands for speed, scale, and ranking accuracy. Table 1 summarises these approaches across embedding granularity, representation style, and retrieval mechanism.

2.4 Question Answering

Question answering systems aim to provide users with direct, contextually appropriate answers rather than requiring them to sift through retrieved documents. Approaches range from factoid-style responses, containing discrete pieces of information, to more complex outputs such as passage extraction or abstractive summaries. This positions QA as a natural progression of information access, aligning more closely with user needs.

Information retrieval systems usually generate a ranked list of documents or re-rank results to directly satisfy a user’s query. To bridge the gap between retrieval and direct answering, many systems introduce a “reader” component

[18, 160], often framed as machine reading [161, 162], or question answering [163–165]. LLMs such as GenQA [166, 167] are increasingly applied to QA tasks. However, they face challenges: hallucinations [168] (e.g. Dahl et al. [169] observed at rates of 69% to 88% in the legal domain), overconfidence despite uncertainty, and limited access to recent or proprietary data [170].

Retrieval-augmented QA. Retrieval augmented generation (RAG) addresses these issues by combining IR with generative models. Retrieved documents ground outputs [168], ensuring factuality, timeliness, and transparency. Some systems further enhance reliability by incorporating references and citations [171]. Some researchers view RAG as a complete system that integrates IR and QA, while others conceptualise QA itself as comprising two components: a retriever, which selects relevant information, and a reader, which generates or extracts the answer [160]. In either view, RAG-based QA represents a subset of broader QA approaches, particularly relevant in open-book scenarios, where grounding in retrieved content enables accurate, context-aware responses.

In cross-lingual QA, methods mirror monolingual IR/QA but often include translation for sparse retrieval or multilingual embeddings for dense retrieval. Research predominantly involves English plus one other language, utilising English’s data abundance to support low-resource settings [98, 172]. Reader modules are particularly beneficial in CLIR systems, as the retrieved documents may be in a language that the user is unable to understand, so the generation of an answer or summary in the original query language allows for the bridging of this language gap.

2.5 Current CLIR Approaches

CLIR has recently drawn from improvements in representation learning, multilingual modelling, and scalable retrieval architectures. Contemporary systems increasingly integrate these elements into full retrieval pipelines, with methods broadly categorised into sparse retrieval, dense retrieval, late-interaction, hybrid, cross-encoders, and multimodal models. In this section, we highlight representative recent work; for definitions and trade-offs of the retrieval architectures mentioned below, refer back to Section 2.3.

Sparse retrieval models. The SPLADE family [147], SPLADEv2 [148], and extensions such as SPLADE-X [151] and MultiSPLADE [173] exemplify this line of work by generating sparse token-level representations and enabling effective multilingual retrieval.

Dense retrieval models. Systems such as LaBSE [174], mSBERT [175], and mDPR [176] demonstrate robust multilingual and zero-shot performance across diverse languages.

Late-interaction models. ColBERT [143] and ColBERT v2 [144] balance fine-grained matching with scalable retrieval, achieving strong effectiveness-efficiency trade-offs.

Hybrid models. Examples include pipelines that merge BM25 with neural retrieval [177, 178], which show effectiveness in low-resource and typologically diverse settings.

Cross-encoder & re-ranking models. Approaches such as Translate-Distill [100], Multilingual RAG [179], and CoConDenser [180] achieve strong performance, while large-scale resources like CLIRMatrix [181] facilitate multilingual evaluation. Recent work, including OPTICAL [182], mContriever-X [183], and SWIM-X [184], further extend retrieval capabilities across dozens of languages.

Multimodal & speech-based CLIR. These systems expand retrieval beyond text. Cross-modal pre-training [185], LECCR [186], and M-SpeechCLIP [187] align text with image or speech embeddings, enabling multilingual retrieval across different modalities.

Other directions. Additional approaches include unsupervised CLIR [188], cross-lingual text encoders [189], and task-specific benchmarks such as CrossMath [190] and MTD/MLIR [191]. Re-rankers increasingly integrate multilingual supervision and teacher-student learning, while hybrid and generative methods adapt retrieval to noisy or adversarial conditions. Despite these advances, gaps remain in domain adaptation, handling morphologically rich and code-switched languages, and incorporating LLMs as multilingual rankers. Addressing these challenges is essential for developing CLIR systems that are accurate, scalable, and equitable.

3 Dealing with Cross-Linguality

While some components of the CLIR pipeline, such as cross-lingual query expansion or translation modules, can be adapted in a modular fashion, approaches specifically developed for cross-lingual retrieval generally yield better performance and more balanced language coverage. Traditional approaches translate queries or documents directly, or use a pivot language, whereas recent methods employ cross-lingual embeddings to map texts into a shared semantic

space for more effective comparison. Alignment strategies such as contrastive learning, adversarial alignment, and self-supervised objectives mitigate linguistic and resource disparities, making them critical for effective and inclusive CLIR systems.

3.1 Translation

Traditional CLIR approaches are usually divided into two main stages: translation and monolingual IR. Unlike full-text translation, CLIR requires only a representation suitable for the retrieval system, meaning that strict syntactic or grammatical fidelity can be relaxed [24]. Queries, which are often very short and ambiguous, are pre-processed through tokenisation, stopword removal, and term expansion [192, 193]. Whereas most translation applications aim to produce a single, readable output, CLIR can benefit from multiple translation alternatives, which can function as part of the query expansion process. After retrieval and ranking, further translation may sometimes be necessary to ensure the user can interpret the documents, but this step is not always required.

Translation Granularity. A central design choice concerns translation granularity: whether to translate the query, the document, or both. Techniques range from dictionaries and traditional Statistical Machine Translation to neural and embedding-based approaches, with pivot or dual translation sometimes needed for low-resource languages. Query translation is more widely used in CLIR than document translation, as it is computationally cheaper, avoids large-scale translation, and can be performed at retrieval time, though it introduces risks of ambiguity and misinterpretation [24, 192, 194]. Document translation, in contrast, is more resource-intensive but benefits from added context and reduced ambiguity. Moreover, the mistranslation of a single word has less effect on retrieval performance [192]. Query translation remains popular due to its efficiency and flexibility, while document translation can be advantageous when all queries are in a single language. Ultimately, query translation may still necessitate subsequent document translation for user access, whereas document translation allows direct examination of retrieved texts.

3.1.1 Translation Techniques

Translation can be classified as “direct” where the source language is translated straight from the target language, or “indirect”, where a pivot language is used to overcome source-target limitations for particular language pairs. Several translation techniques have been developed, including dictionary-based approaches, corpus-driven strategies, statistical models and more recently neural models.

Dictionary-based Methods. These rely on bilingual, machine-readable dictionaries (MRDs) [195, 196]. For each word in the source language, MRDs contain one or multiple synonymous words and phrases in the target language. For each term in the given query, dictionary-based translation simply finds the word in the dictionary and selects the translation. Ambiguity arises because many words have multiple meanings. One solution is to select the most frequent translation [24], while another retains all possible translations in structured query translation [197–199]. The latter improves recall [198] but requires weighting schemes to balance translation probabilities [200–202]. A key limitation of bilingual dictionaries is their poor handling of proper nouns and out-of-vocabulary (OOV) terms, especially newly-coined technical terms [192].

Parallel and Comparable Corpora. Parallel corpora are aligned texts in two languages, such as the Hansard Corpus [203], EuroParl [204], or UN documents [205], and are widely used to induce bilingual dictionaries (e.g. [194, 206]). They support multiple translation options but are costly to collect and often limited in domain. Comparable corpora instead consist of texts that are not translations but share topical or communicative similarity, for example, Wikipedia pages [207–209]. These corpora are easier to obtain, yet translation quality is typically lower than with parallel corpora [192].

Statistical Machine Translation (SMT). SMT was the dominant approach from the 1990s to the early 2010s. It is based on noisy channel models [24, 194], particularly IBM models [210, 211], which assign probabilities to candidate translations and choose the most likely output. A noisy channel model treats the source language text as a misspelled or distorted version of the target language, where the goal is to recover the most likely original target language text [24, 210]. SMT proved effective when large parallel corpora were available, but it typically produced a single best translation, reducing ambiguity that could otherwise aid retrieval [194].

Neural Machine Translation (NMT). NMT has overtaken SMT as the preferred paradigm, replacing phrase-based systems with single neural network architectures [212–214]. Early models handled only one language pair, but subsequent developments expanded to multilingual systems [215]. NMT produces fluent, context-aware translations, handles OOV terms through subword segmentation, and benefits from contextualised embeddings [216, 217]. Encoder-decoder architectures with attention [218–220], and later Transformers [221], enabled more effective long-sequence

translation. Subword methods such as byte pair encoding [222–224], WordPiece [225, 226], SentencePiece [227] are now standard.

NMT has been applied to CLIR through systems such as Translate-Train [228, 229], Translate-Distill [100], mDPR [230] and ColBERT-X [228], which optimise retrieval quality by integrating translation with retrieval. Despite improvements, CLIR performance is still hindered by issues relating to short queries, ambiguity, semantic drift, domain adaptation, and high training data requirements [231]. Overall, while NMT represents a substantial improvement over SMT in terms of fluency, adaptability, and context handling, both approaches remain limited by data availability, computational demands, and scalability challenges in low-resource settings.

3.1.2 Indirect Translation

When direct translation from source to target language is infeasible, indirect translation offers an effective alternative by exploiting resources available for intermediate languages. Two common approaches are transitive translation and dual translation.

Transitive Translation. Transitive translation uses a pivot language: the source text is first translated into a high-resource intermediate language, then into the target (e.g. [232–240]). Gollins and Sanderson [241] highlight triangulation, where the use of multiple pivot languages reduces ambiguity compared to a single pivot, which can accumulate errors. Their findings show that triangulation via three intermediate languages outperforms pairwise merging, though subsequent studies note that its benefits are most evident for unstructured queries [242].

Dual Translation. Dual translation translates both source and target texts into a third language, which can be concrete or abstract (e.g. a semantic space). When concrete languages are used, high-resource ones yield superior translations. Deerwester et al. [243] introduced Latent Semantic Indexing later adapted for CLIR via parallel-corpora [244, 245]. Similarly, Explicit Semantic Analysis (ESA) employs human-readable labels, with Gabrilovich and Markovitch [246] deriving a machine learning approach using weighted vectors. ESA representations, often built from Wikipedia, are interpretable and have been extended to cross-lingual applications.

3.2 Multilingual LLMs for CLIR

The emergence of LLMs, particularly those with multilingual capabilities, has transformed CLIR. Traditional CLIR approaches relied on query or document translation, but the development of transformer architectures [221] and large-scale text corpora has enabled LLMs to achieve strong zero-shot performance with notable generalisation potential. This section reviews standard training steps and architectures for multilingual LLMs. While many LLMs can handle multiple languages, multilingual LLMs are explicitly trained on multilingual corpora. A useful distinction is that if a significant proportion of training data is multilingual, the model can be considered a multilingual LLM [22].

Both LLMs and multilingual LLMs follow three main training stages: pre-training on large corpora, fine-tuning for task specialisation, and alignment with human preferences via reinforcement learning from feedback (RLHF). Architecturally, three variants dominate: encoder-only, encoder-decoder, and decoder-only, with the latter being the most common for text generation. Encoder-decoder models may better suit CLIR [247], particularly for question answering, while decoder-only models are preferable for certain specialised tasks [19].

Different training objectives are used to optimise task-specific performance, evaluated through established metrics and benchmarks. Key challenges include the “curse of multilinguality” [94] discussed below, which highlights trade-offs between performance and multilingual ability. Table 2 summarises widely used multilingual LLMs, with further surveys in [21–23, 248–250].

3.2.1 Training Stages of Multilingual LLMs

As mentioned, the distinction between LLMs and multilingual LLMs is often blurred given that both follow similar training stages, but multilingual LLMs additionally rely on multilingual corpora. Training generally proceeds through pre-training, fine-tuning, and RLHF. Qin et al. [248] highlight the importance of alignment strategies for multilingual performance, distinguishing parameter-tuning alignment from parameter-frozen alignment, where alignment occurs after post-training via methods such as prompting or code-switching.

Pre-training is based on large-scale multilingual corpora. These usually consist of monolingual texts across many languages (still predominantly English) and a smaller set of parallel corpora, which are less widely available [23, 249]. Typical sources include Common Crawl and Wikipedia. The aim is knowledge acquisition and learning universal language structures. Pre-training can start from scratch, with parameters randomly initialised, or follow a continual pre-

Model	Architecture	Languages	Open-source	Training Data
mBERT [251]	E	104	✓	Wikipedia
XLNet [94]	E	100	✓	CommonCrawl
mContriever [183]	E	29	✓	CCNet and Wikipedia
ColBERT-X [228]	E	7	✓	MS MARCO
Qwen3-Reranker [252]	E	100+	✓	Multilingual query-document pairs
multilingual-E5-large [253]	E	100+	✓	1B text pairs and retrieval/TyDi tasks
XLNet-V [254]	E	100+	✓	CommonCrawl
Nomic Embed v2 [255]	E	100+	✓	Data from a variety of sources
Gemini Embedding [256]	E	100+	✗	Undisclosed
LaBSE [257]	E	109+	✓	CommonCrawl, Wikipedia and webpage translation pairs
mT5 [95]	E-D	101	✓	CommonCrawl
mBART [258]	E-D	25	✓	CommonCrawl
NLLB-200 [259]	E-D	202	✓	Parallel data
mLongT5 [260]	E-D	101	✓	mC4
Qwen3-Omni	E-D	119	✓	Undisclosed
Aya [261]	E-D	101	✓	xP3x, Aya Dataset, Aya Collection, Data Provenance and ShareGPT-Command
XGLM [262]	D	30	✓	CommonCrawl
LLaMA 3 [263]	D	30	✓	Publicly available data
Mistral [264]	D	Dozens	✓	Publicly available data
BLOOM [265]	D	46 natural, 13 programming	✓	ROOTS
Yi-01 [266]	D	2 (English and Chinese)	✓	Publicly available data including CommonCrawl
GPT-4o [267]	D	50+	✗	Publicly available and proprietary data
Gemini 2.5 [268]	D	ND (40+)	✗	Publicly available data
Claude 3 [269]	D	ND	✗	Publicly available and proprietary data
PaLM 2 [270]	D	ND (100+)	✗	Data from a variety of sources including Wikipedia, webpages and news articles
DeepSeek-V3 [271]	D	ND	✓	Data from a variety of sources
Gemma [272]	D	20+	✓	Data from a variety of sources
PolyLM [273]	D	18	✓	mC4, CC-100, The Pile, GitHub and OPUS
Nemotron-4 15B [274]	D	53+	✓	Data from a variety of sources (53 languages, 43 code)

Table 2: Overview of explicitly multilingual LLMs as well as LLMs with multilingual capabilities. Architecture: E = encoder-only, E-D = encoder-decoder, D = decoder-only. ND = not disclosed.

training approach, where an existing LLM is adapted with multilingual or domain-specific data [21, 22]. While continual pre-training is computationally cheaper and allows faster domain adaptation, it risks “catastrophic forgetting” [275] where previously learned knowledge is lost due to distributional shifts between old and new data [276]. To mitigate this, methods such as replay buffers, parameter freezing, and elastic weight consolidation are used [277].

Supervised fine-tuning (SFT) adapts pre-trained models to specific tasks using labelled datasets. Unlike pre-training, which focuses on broad knowledge acquisition, fine-tuning specialises models for instruction following or task-specific objectives [21]. Increasingly, fine-tuning data combines human- and model-generated content, with effectiveness depending on data quality and diversity. Beyond instruction tuning, multilingual SFT extends to tasks such as CLIR, Named Entity Recognition (NER), Sentiment Analysis and Text Classification [202].

RLHF further aligns multilingual LLM outputs with human preferences. Human annotators either rank outputs or select between alternatives [278, 279], producing preference data used to train a reward model. The base model is then fine-tuned using algorithms such as Proximal Policy Optimisation [280]. RLHF is resource-intensive due to the need for large-scale human annotation, though synthetic data has reduced some costs. Nonetheless, concerns remain over the potential for manipulative behaviours to be learned by models when optimising for human feedback [281].

3.2.2 Model Architectures of Multilingual LLMs

Like monolingual LLMs, multilingual LLMs are based on the transformer architecture [221], which consists of encoder and decoder modules that rely on self-attention. Variants fall into three categories: encoder-only, encoder-decoder,

and decoder-only (see Figure 3 for an illustration and Table 2 for details on the architectures of the listed multilingual LLMs).

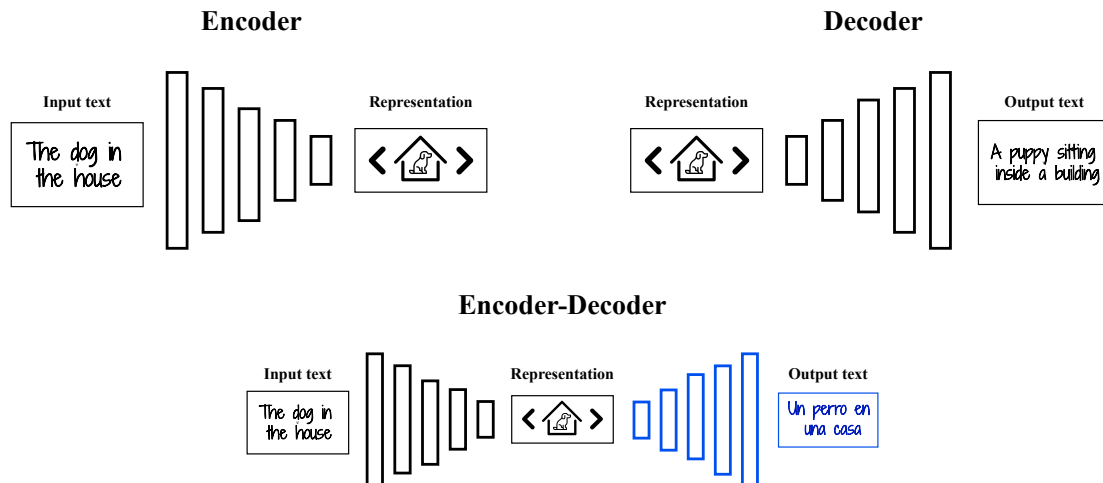


Figure 3: LLM model architectures.

Encoder-only models (e.g. BERT [251]) tokenise input into embeddings that encode semantic and positional information. These embeddings are processed by stacked encoder layers with multi-head self-attention and feed-forward networks, producing contextualised representations of the input [249]. Because they attend bidirectionally, encoder-only models excel at transforming data into compressed representations useful for language understanding tasks such as sentiment analysis, NER, and classification in both monolingual and multilingual settings [23]. However, they are not well suited for generative tasks like next-token prediction or translation.

BERT was initially released as a monolingual English model, which then led to other language-specific variations (e.g. FlauBERT [282] for French, BERTje [283] for Dutch and AfriBERT [284] for Afrikaans), eventually resulting in the multilingual version mBERT that is trained on 104 languages. Other multilingual encoder-only models include XLM-R [94] and LaBSE [257].

Encoder-decoder models preserve both components of the transformer, allowing the encoder to process input and the decoder to generate output conditioned on the encoded representation [221]. This makes them well-suited to sequence-to-sequence tasks such as summarisation and, in the multilingual domain, machine translation. Examples include mT5 [95] and mBART [258].

Decoder-only models omit the encoder and generate text autoregressively, producing tokens sequentially while attending only to previous tokens. This unidirectional structure makes them effective for text generation and completion tasks [21, 23]. Their popularity has grown significantly with the release of GPT-style models. Widely used multilingual decoder-only models include GPT-4o [267], PaLM [285], XLGM [262] and BLOOM [265].

3.2.3 Multilingual Pre-training Datasets

Monolingual LLMs are trained on large monolingual corpora. Multilingual LLMs follow a similar approach, extending the corpora to monolingual texts in multiple languages and parallel corpora (collections of translations). Model performance depends on corpus choice: some prioritise higher-resource languages, while others (e.g. IndicBERT [286] for 12 Indian languages, or AfriBERTa [287] for 11 African languages) target low-resource settings. Broader multilingual coverage requires varied corpora, including mid- and low-resource languages.

Monolingual Corpora. Massive monolingual corpora enable learning universal language representations, critical for multilingual models, and reduce reliance on parallel data. Most pre-training data comes from web sources, particularly Common Crawl and Wikipedia. Crawled data, however, often contains harmful or low-quality content; thus, cleaned and filtered datasets are used (e.g. CC-100 for XLM-R [94]), and additional training stages like RLHF aim to mitigate undesirable outputs. Despite this, English dominates many training corpora. For example, Xu et al. [23] report that English comprises about 92.1% of ChatGPT’s training corpus, leaving relatively little representation for other widely spoken languages. Xu et al. [23] further highlight that when English is excluded, Indo-European languages (e.g. German, French) still constitute over 50% of the remaining data in their language-family analysis.

Parallel Corpora Parallel corpora in multilingual pre-training resemble those in machine translation models, including manually created datasets (e.g. Bible Corpus [288], MultiUN [289]) and machine-generated corpora via multilingual LLM-aided generation.

3.2.4 Multilingual LLM Training Objectives

Pre-training objectives for LLMs and multilingual LLMs aim to specialise models for specific tasks. Doddapaneni et al. [250] categorise them into three types. The first adapts monolingual functions for multilingual use, e.g. Probabilistic Language Modelling [290], Masked Language Modelling [251] and Next Sentence Prediction [251]. The second leverages parallel corpora at the sentence/document level, including Translation Language Modelling [291], Cross-Attention Masked Language Modelling [292] and Cross-Lingual Masked Language Modelling [293]. The third exploits other parallel resources such as word alignments, e.g. Cross-Lingual Word Recovery [293], Alternating Language Model [294] and Back Translation Masked Language Modelling [292]. Table 3 presents the most common objective functions used to train multilingual LLMs following this categorisation [250], and Table 4 provides illustrative examples of some of these multilingual training objectives using the sentence pair “A dog in a house” and “Un perro en una casa”.

3.2.5 The Curse of Multilinguality

Conneau et al. [94] describe the “curse of multilinguality,” where the inclusion of additional languages during pre-training improves performance up to a point, after which both monolingual and cross-lingual performance declines. This reflects a trade-off between expanding language coverage and maintaining per-language capability. Proposed solutions include parameter sharing across pre-training languages such as Cross-lingual Modular [302], or language-specific parameter subsets such as those proposed by Blevins et al. [303] for Cross-lingual Expert Language Models (X-ELM). Another approach is that employed by Artetxe et al. [304], who train with a masked objective in one language before extending to a new language via an embedding matrix that freezes earlier parameters, ensuring stability while enabling transfer to new languages.

3.3 Embeddings

Transformer-based models have become central to information retrieval due to their ability to produce contextualised text embeddings. Unlike co-occurrence-based representations, embeddings encode deep semantic information, enabling cross-lingual comparison without explicit translation. Early work relied on static embeddings and alignment techniques, supervised [305] or unsupervised [306], which enabled bilingual lexicon induction and translation by mapping monolingual embeddings into a shared space.

The introduction of textual models such as BERT [251] marked a major shift. Multilingual variants like mBERT [251] extended pre-training to dozens of languages, showing promising zero-shot transfer, though scaling to many languages often reduced performance due to the aforementioned curse of multilinguality [94, 307]. Models such as XLM [291] addressed this by incorporating translation-based objectives like Translation Language Modelling (i.e. leveraging sentence-level parallel data to strengthen cross-lingual transfer), while RoBERTa [308] introduced architectural improvements, leading to XLM-RoBERTa [94] as a strong multilingual baseline.

These models are particularly impactful for low-resource languages, supporting zero-shot [309] and few-shot [310, 311] transfer, and enabling applications with limited data [312]. Multilingual embeddings have proven effective for cross-lingual retrieval, surpassing translation-based methods [313, 314]. By embedding text from multiple languages into a shared space, they allow direct comparison of meaning across languages [12, 315] without the need for translation. In CLIR, this enables retrieval of semantically relevant documents across languages based on embedding similarity rather than parallel corpora or keyword overlap [257, 316].

Creating Multilingual Embeddings. The most common approach to constructing multilingual embeddings is *joint pre-training*, where a model is trained on multilingual corpora using Masked Language Modelling [20]. This yields language-independent representations by exploiting shared structural and lexical regularities. The quality of embeddings depends on training data, with alignment enhanced via parallel corpora at the word [317], sentence [318], or domain [319] level. Large bilingual dictionaries further improve semantic alignment [320]. The aim is to map semantically equivalent concepts to nearby positions in the embedding space, as visualised in Figure 4.

Before contextual embeddings, cross-lingual retrieval relied on mapping functions aligning monolingual word embeddings. Early work applied linear transformations from bilingual dictionaries [305], while unsupervised methods such as MUSE [321] used adversarial learning. Later models refined alignment iteratively [322], enabling applications in lexicon induction and retrieval, but remained limited by domain differences and pairwise mapping requirements [323]. This motivated the shift toward multilingual pre-training and contextual embedding spaces.

Training Objective	Description
Adapted from monolingual	
Probabilistic Language Modelling (PLM) [290]	Estimates the probability distribution of sequences of words in a language.
Masked Language Modelling (MLM) [251]	Inspired by a Cloze task [295]. Certain tokens are randomly masked, and the model predicts the masked tokens based on the available context. This objective encourages the model to learn bidirectional representations and dependencies between words in a sentence.
Next Sentence Prediction (NSP) [251]	Predict whether a given pair of sentences is contiguous or not. Through this objective the model learns to understand coherence and logical flow between sentences.
Denoising Autoencoder (DAE) [296]	Given a partially corrupted or noisy input, the model is trained to recover the original undistorted input.
Causal Language Modelling (CLM) [291]	Autoregressive next-token prediction: predict the next token in a sequence of tokens; the model has access to unidirectional context.
Multilingual Replaced Token Detection (MRTD) [297]	Tokens are replaced in a multilingual sequence, and the model is trained to detect which are the real input tokens from the corrupted sentences.
Parallel corpora	
Translation Language Modelling (TLM) [291]	Sentences in different languages are concatenated, and tokens are masked at random. The model then has to predict the masked tokens.
Cross-Attention Masked Language Modelling (CAMLML) [292]	Using a parallel sentence pair, the model is trained to predict masked tokens in one language using the other language.
Cross-Lingual Masked Language Modelling (CLMLM) [293]	Similar to TLM, but the input is constructed at the document level. Sentences in a cross-lingual document are masked at random, and the model is trained to predict these masked tokens.
Cross-Lingual Contrastive Learning (XLCO) [298]	Contrastive learning is used: the model learns to bring representations of semantically similar sentences together, and push negative pairs apart.
Hierarchical Contrastive Learning (HICTL) [299]	Contrastive learning is applied both at sentence and word-level, with the goal for the model to learn language-invariant sentence representations.
Cross-Lingual Sentence Alignment (CLSA) [300]	Using parallel data, the model is encouraged to align sentence representations across languages.
Translation Replaced Token Detection (TRTD) [297]	Given translation pairs, the model is trained to detect masked tokens in both languages.
Parallel resources	
Cross-Lingual Word Recovery (CLWR) [293]	Goal is to learn underlying word alignments between two languages by predicting missing words in one language using aligned source sentences.
Cross-Lingual Paraphrase Classification (CLPC) [293]	Given two sentences from different languages, classifies whether they have the same meaning.
Alternating Language Model (ALM) [294]	Using code-switched sentences (alternating languages between phrases), the model is trained to predict masked language modelling.
Denoising Word Alignment (DWA) with Self-Labeling [301]	Two alternating steps: word alignments are first estimated, and the model then predicts masked tokens in parallel sentence pairs.
Bidirectional Word Alignment (BWA) [300]	Employs the attention mechanism in the transformer model to align word representations across languages using parallel data.
Back Translation Masked Language Modeling (BTMLM) [292]	Tokens from a source language are predicted (translated) into a target language. Tokens in the source language are then masked, and the target language tokens are used to predict them.

Table 3: Multilingual LLM training objectives. The objectives are classified into those that are adapted from monolingual training objectives, those that leverage parallel corpora at the sentence or document level, and those that exploit other parallel resources such as word alignments.

Even with shared spaces, mismatches often occur due to style or structural differences, weakening similarity measures. To address this, mapping functions project queries and documents into shared spaces using methods such as student-teacher training [324], geometric alignment [325], or post-hoc projections [326]. These lightweight techniques provide efficient alternative to full fine-tuning or translation-based retrieval.

Monolingual-Inspired Objectives	
Masked Language Modelling	EN: A dog in a [MASK]. Context: “dog”, “in” → predict “house”
Causal Language Modelling	EN: A → dog → in → a → house → [EOS] Predict next token based on previous tokens only
Multilingual Replaced Token Detection	ES: Un perro en una casa. Detect whether “una” is real or replaced (binary label)
Parallel Corpora Objectives	
Translation Language Modelling	EN: A dog in a [MASK]. ES: Un [MASK] en una casa. Predict “house” and “perro” using cross-lingual context
Cross-Attention MLM	EN: A dog in a house. ↑ ES: Un perro en una [MASK]. Use EN tokens via cross-attention to predict “casa”
Cross-Lingual Contrastive Learning	Positive pair: EN: A dog in a house. →← ES: Un perro en una casa. <i>Pull their embeddings closer</i> Negative pair: EN: A dog in a house. ←→ ES: El gato duerme. <i>Push these embeddings apart</i>
Parallel Resources Objectives	
Cross-Lingual Word Recovery	EN: A dog in a [MASK]. ↑ Aligned: ES: Un perro en una casa. Use aligned “casa” to recover “house” (alignment rather than attention, unlike Cross-Attention MLM)
Alternating Language Modelling	Mixed: A perro in una [MASK]. Predict “casa” using context in both languages
Bidirectional Word Alignment	EN: A dog in a house. ↑ ↑ ↑ ↑ ↑ ES: Un perro en una casa. Model learns token-to-token alignments in both directions

Table 4: Illustrative examples of multilingual training objectives using the sentence pair {EN: *A dog in a house*, ES: *Un perro en una casa*}, following the same classification as Table 3.

3.4 Alignment in CLIR

Effective alignment of embedding spaces is central to CLIR. It ensures semantically equivalent content across languages, whether this be queries, documents, or other distributions, is mapped to comparable vector representations. This enables retrieval without explicit translation, parallel corpora, or shared vocabulary. Cross-lingual alignment positions semantically similar words, phrases, or sentences from different languages close in the embedding space (Figure 4), ideally abstracting away surface-level differences (e.g. syntax, orthography) in favour of shared meaning.

Alignment is particularly valuable as it allows models trained in one language to generalise across others. Even fine-tuning on monolingual query-document pairs can update monolingual embedding spaces, enabling transfer across languages and reducing the need for language-specific supervision, which is especially critical for low-resource settings [175, 189, 327, 328].

However, alignment does not naturally arise from multilingual pre-training alone. Models trained on large multilingual corpora often develop language-specific subspaces. Thus, additional mechanisms are required to ensure semantically equivalent content aligns across languages.

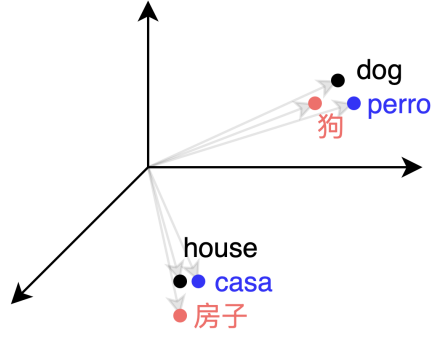


Figure 4: Simplified representation of a multilingual embedding space, highlighting how semantically equivalent concepts are mapped to nearby positions regardless of language.

3.4.1 Levels and Methods of Alignment

Alignment can be achieved at different granularities. Word-level alignment maps individual lexical tokens but is limited by lexical gaps and cultural specificity [20, 321]. Sentence-level alignment instead captures entire sentences or phrases, preserving semantic meaning while mitigating lexical mismatches. Tools like Awesome-Align [329] and SimAlign [330] exploit parallel corpora, while parallel-data objectives such as TLM in XLM [291], where aligned sentences are concatenated and masked so the model learns from both monolingual and cross-lingual context, or LASER, [327] enhance alignment through encoder-based architectures.

Contrastive learning has emerged as a scalable alternative when parallel data are scarce. It optimises directly over paired inputs: semantically equivalent pairs are pulled closer, while dissimilar ones are pushed apart. Applied to CLIR, this often involves query-document pairs (positive/negative) generated via translation or pseudo-parallel data [180, 331]. Objectives such as triplet loss [332], InfoNCE [333], NT-Xent [334], and Multiple Negatives Ranking Loss [335, 336] improve efficiency and scalability, making them well-suited for CLIR. Models like LaBSE [174] and mSBERT [175] apply contrastive losses for multilingual sentence embeddings, achieving zero-shot transfer. ALIGN [337] further scales contrastive learning to massive multilingual datasets, demonstrating generalisation across modalities and languages, while CoConDenser [180] combines contrastive pre-training with dense retriever fine-tuning for stronger retrieval benchmarks.

Generative Adversarial Networks (GANs) have also been used to map source embeddings into target-language spaces [338]. Early methods like MUSE [321] showed unsupervised adversarial alignment could rival supervised baselines. Yet GAN-based methods remain unstable and less effective in CLIR due to difficulties in handling polysemy (see for example English–Spanish polysemy mismatch illustrated in Figure 5), compositional semantics, and high-dimensional instability [323]. These methods typically assume that the source and target embedding spaces are approximately isomorphic, an assumption that holds reasonably well for similar languages with comparable corpora and training objectives, but often fails in practice for distant languages or mismatched domains, leading to degraded alignment quality and limited transfer effectiveness [323].

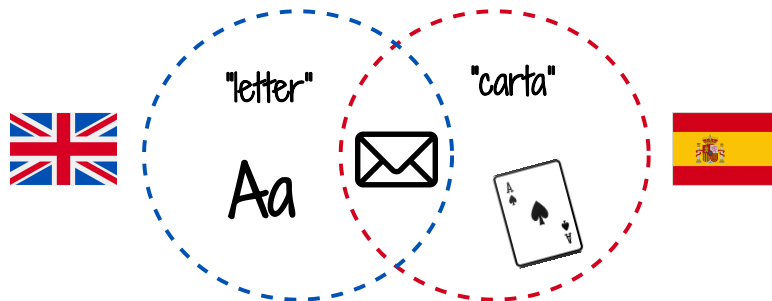


Figure 5: Cross-lingual polysemy mismatch example using the word pair {EN: *letter*, ES: *carta*}.

3.4.2 Query-Document Alignment

Unlike cross-lingual alignment, aligning queries with documents is complicated by their inherent asymmetry. Queries are short, focused, and varied, while documents are longer, diverse, and often cover multiple topics. A single document may serve numerous queries, but a query typically has only a few relevant documents. Consequently, strict embedding alignment fails to capture the necessary contextual nuances. To address this, dual-encoder models use separate query and document encoders while training to align them in a shared space. Dense Passage Retrieval [339] exemplifies this by jointly encoding queries and answers, while later models like ANCE [340] and TAS-B [341] improve via dynamic negative sampling. These approaches aim to optimise embeddings for semantic similarity rather than enforcing exact pairwise alignment.

A central issue in query-document alignment is whether to: (i) map queries into the document embedding space (the more common approach) - queries are often expanded or transformed to better resemble documents, thereby reducing style and length mismatches (see Section 2.1); or (ii) map documents into the query embedding space. The latter is less common, but useful when aiming to directly align document semantics with short, focused queries.

Through LLM-based generation, this idea extends to producing pseudo-documents from queries or pseudo-queries from documents (see Section 3.2). In both directions, the objective is to minimise the gap between generated text and its true match. Another promising direction is fine-tuning embeddings of generated text so they align directly with those of their correct counterparts.

3.4.3 Is Alignment Necessary?

While cross-lingual alignment is a desirable property, it is neither always necessary nor sufficient for transfer of model capabilities in tasks like CLIR. Fine-tuning may even weaken initial alignment [342], and alignment itself can reinforce cultural or linguistic biases inherited from high-resource languages [343]. Instead, CLIR performance often benefits more from representations that preserve semantic distinctions and disambiguate meaning in context. Effective retrieval depends on resolving polysemy, retaining query-relevant information, and ensuring embeddings capture objective-specific semantics, even if perfectly co-located alignment is absent [344, 345].

Thus, what ultimately matters for CLIR is not strict geometric uniformity but whether embedding proximity reflects semantic relevance. Approximate, task-sensitive alignment combined with contextual understanding often proves more valuable than perfect alignment.

4 Evaluation

Since CLIR systems are composed of multiple components, evaluation can be conducted either component-wise or end-to-end. Full-system evaluation is essential for assessing performance on the overall task of cross-lingual retrieval [346], spanning RAG, QA, and domain-specific retrieval. Evaluation employs diverse datasets and metrics, including those from machine translation, monolingual IR, multilingual generation, and embedding alignment. While some datasets provide gold labels for cross-lingual relevance, many rely on proxy or reference-free metrics. Metrics may target lexical overlap, semantic similarity, or ranking quality, depending on the evaluation goal. A range of CLIR benchmarks support this, covering multiple retrieval settings. This section reviews the most relevant datasets and evaluation practices for CLIR, spanning both system-level and component-level assessment.

4.1 Translation Evaluation

In CLIR, translation is not assessed for readability or fluency, but for retrieval effectiveness. Thus, the focus shifts from translation accuracy to its impact on retrieval. Traditional metrics like Bilingual Evaluation Understudy (BLEU)², which evaluates n-gram overlap, remain widely used but have limitations, such as poor handling of synonym and word order. Alternatives include the Metric for Evaluation of Translation with Explicit Ordering (METEOR)³ score, which correlates better with human judgements, and Translation Edit Rate (TER)⁴, which measures edit distance to reference translations.

²<https://huggingface.co/spaces/evaluate-metric/bleu>

³<https://huggingface.co/spaces/evaluate-metric/meteor>

⁴<https://huggingface.co/spaces/evaluate-metric/ter>

Recent neural metrics leverage pre-trained language models, such as BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)⁵ and COMET (Crosslingual Optimized Metric for Evaluation of Translation)⁶. These better capture semantic adequacy and fluency. For retrieval-oriented evaluation, recall is generally prioritised over precision, as the goal is to ensure relevant documents are retrieved, minimising false negatives. Although handling irrelevant documents remains important, precision is usually secondary in CLIR [25].

4.2 Datasets for CLIR

A major challenge in CLIR is dataset availability, particularly for low-resource languages. Table 5 lists multilingual datasets for CLIR and cross-lingual QA. Large-scale CLIR datasets covering diverse topologies include CLIRMatrix [181], Large-scale CLIR Dataset [347], SWIM-IR [184], XOR-TyDi QA [98], LAReQA [348] and mMARCO [97]. Other datasets focus on specific languages (e.g. WikiCLIR [349], NeuMARCO [350], CLIRudit [351]), or regions (e.g. AfriCLIRMatrix [178], CIRAL [352]). BordIRlines [353] targets geopolitical disputes, while MIRACL [99] and other monolingual IR datasets contain diverse languages but are not CLIR-specific. Cross-lingual resources include Mr. TyDi [230], XQA [354], MLQA [113], XQuAD [304], TyDi QA [355], MKQA [356], and XRAG [357].

Most CLIR datasets are derived from Wikipedia, which offers a broad coverage but limited domain diversity and data variation. As with LLM training data, English dominates, often serving as the query or document language (e.g. AfriCLIRMatrix [178], NeuMARCO [350], WikiCLIR [349], XOR-TyDi QA [98]). This reliance highlights the cross-lingual imbalance, with English functioning as the basis for transfer to lower-resource languages.

Many multilingual datasets extend monolingual English IR/QA corpora, such as TyDi QA [355] and MS MARCO [358]. The latter is notable as it is created using human-generated text, whereas most others rely on automatically created content. Increasingly, datasets covering additional languages use LLM-aided generation and machine translation, producing query-document triplets (query, positive, negative). These methods support query expansion, training augmentation, and evaluation particularly valuable for low-resource languages. Finally, large-scale multilingual parallel corpora (e.g. OPUS⁷) are widely used to train MT systems and LLMs, though they fall outside strict CLIR scope.

4.3 CLIR Evaluation Metrics

Evaluating CLIR systems typically relies on the Cranfield paradigm, which defines a test collection by fixing a document corpus, a set of query topics, and relevance judgments, enabling reproducible, comparable assessment of retrieval models [360]. Recall often takes precedence over precision or F_1 scores, as identifying all relevant documents is more critical than excluding irrelevant ones. Rank-aware measures such as Mean Reciprocal Rank (MRR) [361] and nDCG [91] build on recall/precision to reward early retrieval of relevant results [362]. CLIR adds further evaluation complexities: resource-imbalances and over-looked low-resource languages introduce potential biases, while translation or embedding quality can significantly affect retrieval performance. Although standard IR metrics remain central for end-to-end evaluation, complementary measures are also used, such as translation fluency scores, back-translation retrieval, or embedding alignment, even if they are not typically employed as the final evaluation criteria.

4.3.1 Retrieval Metrics

Among retrieval metrics, Hit Ratio@K (Hit@K) [361] measures whether at least one relevant document appears in the top K results for a query, focusing on binary relevance. Recall@K [92] computes the fraction of all relevant documents returned in the top K , thus highlighting completeness. MRR [361] calculates the inverse rank of the first relevant document for each query and averages across queries, rewarding early retrieval. MAP [92] averages precision scores across all relevant documents and queries, favouring systems that rank relevant documents earlier. Discounted Cumulative Gain (DCG) [91] assigns higher weights to relevant documents retrieved earlier, using a logarithmic discount. Its normalised variant, nDCG [91], divides DCG by the ideal DCG, allowing fair comparison across queries with different numbers of relevant documents.

MRR and MAP are especially useful when early retrieval matters, while nDCG is popular in multilingual and graded-relevance contexts (e.g. cross-lingual QA), as it considers both rank position and degree of relevance. In recent years, nDCG and MRR have emerged as preferred metrics for evaluating dense and neural retrieval systems due to their robustness in handling ranking nuances and graded relevance judgements.

⁵<https://huggingface.co/spaces/evaluate-metric/bleurt>

⁶<https://huggingface.co/spaces/evaluate-metric/comet>

⁷<https://opus.nlpl.eu>

Dataset	Application	Description	Data source
CLIRMatrix [181]	CLIR	Two IR datasets: bilingual dataset in 139 languages (BI-139) and multilingual dataset in 8 languages (MULTI-8)	Wikipedia
AfriCLIRMatrix [178]	CLIR	English queries with relevance judgements for documents in 15 African languages	Wikipedia
Large-Scale CLIR Dataset [347]	CLIR	English queries, relevant documents in 25 other languages	Wikipedia
SWIM-IR [184]	CLIR	Query-passage pairs in 33 languages, queries generated by PaLM-2 from Wikipedia passages using summarise-then-ask prompting	Wikipedia
CIRAL [352]	CLIR	English queries, documents in four African languages (Hausa, Swahili, Somali and Yoruba)	News articles
WikiCLIR [349]	CLIR	German queries and English documents	Wikipedia
XOR-TyDi QA [98]	CLIR	Expanded TyDi QA dataset into 7 typologically diverse languages, documents in English	Wikipedia
LARQA [348]	CLIR	Convert XQuAD and MLQA into answer retrieval tasks (XQuAD-R and MLQA-R) in 11 languages	Wikipedia
NeuMARCO [350]	CLIR	English queries with documents from MS MARCO translated into Chinese, Persian, and Russian (machine translated)	Human generated
mMARCO [97]	CLIR	Multilingual version of MS MARCO in 13 languages (machine translated)	Human generated
CLIRudit [351]	CLIR	French and English research articles with English queries and French documents	Érudit
BordIRlines [359]	CL-RAG	Queries about geopolitical disputes, documents in languages covering all claimant countries for each territory (251 disputes, 720 queries, 49 languages)	Wikipedia
MIRACL [99]	IR	Relevance judgements for documents in 18 languages from 10 language families for monolingual retrieval	Wikipedia
Mr. TyDi [230]	IR	Question-passage pairs 11 typologically diverse languages for monolingual retrieval	Wikipedia
XQA [354]	CL-QA	Questions, answers and top 10 retrieved articles in 9 languages	Wikipedia
MLQA [113]	CL-QA	7 languages (English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese), each instance parallel between 4 languages on average	Wikipedia
XQuAD [304]	CL-QA	Translated subset of SQuAD v1.1 into ten languages (Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi)	Wikipedia
TyDi QA [355]	CL-QA	Question-answer pairs in 11 typologically diverse languages	Wikipedia
MKQA [356]	CL-QA	Question-answer pairs in 26 typologically diverse languages	Web/human generated
XRAG [357]	CL-RAG	Multilingual RAG across 5 languages, supporting both document retrieval and generation tasks; includes natural queries and Wikipedia passages	Wikipedia

Table 5: Multilingual datasets for information retrieval and question answering.

4.3.2 Cross-lingual Performance Metrics

Measuring the cross-lingual ability of language models introduces challenges as models must handle different scripts, linguistic structures, and cultural contexts. General multilingual benchmarks such as M-RewardBench [363], INCLUDE [364], and Global-MMLU [365] address these aspects while differing in task format but sharing a common evaluation metric of accuracy. M-RewardBench evaluates reward models across 23 languages by testing whether they correctly prefer human-preferred responses in paired comparisons spanning chat, safety, reasoning, and translation. INCLUDE tests knowledge and reasoning using about 197k multiple-choice questions sourced from local exams in 44 languages,

scoring models by the fraction of correct answers. Global-MMLU extends the original MMLU into 42 languages with professional translation and annotation, using multiple-choice accuracy to assess model understanding and offering additional analysis on culturally sensitive versus agnostic subsets.

Beyond benchmark accuracy, specialised metrics assess multilingual models at finer levels, with and without gold-standard labels. Translation and generation metrics include BLEU [366], which as mentioned above measures n-gram precision between system output and reference with a brevity penalty; ROUGE [367], which emphasises recall of n-grams, subsequences, and skip-bigrams often used in summarisation; METEOR [368], which incorporates stemming, synonym matching, and alignment-based precision and recall for stronger correlation with human judgements; chrF [369], a character level F-score suitable for morphologically rich languages; and TER [370], which computes the number of edits required to change a hypothesis into a reference.

Embedding alignment metrics target semantic coherence across languages. Backretrieval [371] measures alignment quality by checking whether multilingual captions retrieve the same image, while MEXA [372] evaluates how well non-English sentences align with English-centric representations in multilingual models. Learned metrics rely on pre-trained encoders to approximate human judgements. BERTScore [373] computes token-level semantic similarity using contextual embeddings from BERT, COMET [374] trains a regression model on pre-trained encoders to predict human judgement scores, and BLEURT [375] fine-tunes BERT to estimate human-likeness scores based on reference comparisons.

In practice, standalone cross-lingual metrics are rarely applied directly in CLIR evaluation beyond development and diagnostic stages. Most deployed CLIR methods depend on standard retrieval metrics applied to multilingual benchmarks. Translation-quality metrics often inform CLIR evaluation by serving as extrinsic measures of query or document translation quality. Alignment-focused metrics are also used for probing multilingual embedding coherence or representation quality, typically in ablation studies or auxiliary evaluations with pivots such as English. Toolkit-based proxies such as CLIReval [376] attempt to bridge the gap by adapting machine translation evaluation datasets into retrieval tasks, binding translation-oriented metrics to retrieval performance. Overall, while cross-lingual metrics support model development, alignment analysis, and translation-tuning, CLIR evaluation remains primarily centred on retrieval-focused performance.

4.3.3 Generation Evaluation Metrics

For question answering, model performance is often evaluated through human judgments, typically using crowdworkers, though they may lack the expertise to assess factuality and other qualities accurately [377, 378]. A/B testing is common, with annotators comparing answers (e.g. HURDLES [379], WEBGPT [380]), while some studies instead rely on domain experts [381], though they are harder to obtain. Human annotators typically emphasise attributes such as relevance, factuality, and ease of understanding, which are difficult to capture automatically, meaning there is no single comprehensive metric that does not require human annotators. Existing automatic metrics, often adapted from summarisation tasks rather than designed for QA, include measures like ROUGE [367], BERTScore [373] and BLEURT [375], which require human-written references and are limited for long-form QA due to the diversity of valid answers [379, 382]. Automatic metrics frequently fail to align with human judgments [381], tending to capture narrow aspects such as fluency or query relevance, while factuality remains especially difficult, with some approaches borrowing from summarisation faithfulness metrics such as QAFactEval [383]. Additional metrics include Self-BLEU [384], which measures the diversity of generated text, and Perplexity [385], which evaluated linguistic fluency. Particularly pertinent to CLIR, some metrics instead capture the relevance of a question to a given answer (e.g. RankGen [386], BARTScore [387] and Question Likelihood [388]) - these can be useful for QA in information retrieval.

5 Applications

CLIR is essential whenever queries, documents, or both queries and documents appear in multiple languages. Since English dominates online content (English makes up 49.1% of websites as of February 2025, followed by Spanish (6.0%), German (5.6%), and Japanese (5.0%) [389], see Figure 6), users of low-resource languages face significant barriers [390]. CLIR addresses this imbalance by enabling access to information beyond a user’s native language.

Search Engines. To provide relevant results, search engines must retrieve content in higher-resource languages and present it in accessible ways [391]. Cross-lingual search on Google underperforms compared to monolingual search [391], though syntactic analysis improves results [392]. Tools such as Google Translate and AI summaries for search engines highlight the potential of CLIR for translation and summarisation [393].

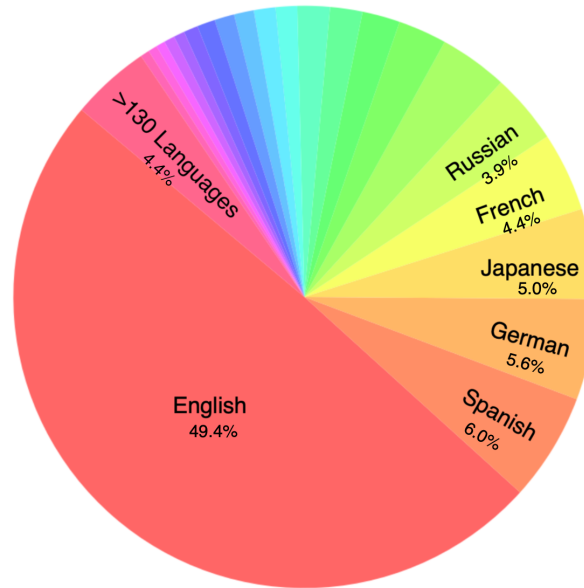


Figure 6: Share of websites by content language as of February 2025. Data represents the most frequently used languages for web content, based on analysis published by Statista [389].

Specialised Databases. Professionals in fields such as law and medicine rely on technical resources often available only in English. CLIR enables access without requiring duplicate content in multiple linguistic versions. Examples include English–Persian retrieval [394] and Hindi–English medical retrieval using morphology and query expansion [395].

LLMs and Question Answering. Retrieval-augmented generation systems benefit from CLIR by retrieving relevant information from multilingual sources and generating answers independent the source language. This makes them effective in multilingual environments where queries and evidence span different languages [179]. At the same time, benchmarks point to challenges in cross-lingual QA and reasoning [353, 357], underlining areas where further progress is needed.

News, Media and Security. Journalists, researchers, and businesses use CLIR to monitor global events and gather multilingual insights. Benchmarks such as NeuCLIR [350], xMIND [396], and CLSD [397] evaluate news retrieval, while MMTweets [398] and MultiClaim [399] assess fact-checking. These functions are equally important in crisis response and security, where timely access to local protocols, reports, and threat indicators supports situational awareness and the detection of terrorism and cyber activity [400–408].

Scientific Research. English dominance in academic publishing creates obstacles for non-English-speaking researchers. CLIR enables access to international literature without requiring translation into English, fostering collaborations and reducing duplications. Benchmarks such as CLIRudit [351] and OPTICAL [182] address retrieval over academic documents, while studies track misinformation diffusion and evaluate claim verification across languages.

E-commerce. E-commerce platforms apply CLIR to help customers search product catalogues and reviews in their native language. Benchmarks such as CLPR-9M [409] and multilingual ranking systems [84] show improvements in retrieval, while datasets like xPQA [410] support cross-lingual product question and answering.

Across domains, CLIR reduces language barriers and ensures equitable access to information. It enables accurate retrieval in medicine, law, research, media, security, and commerce, while promoting inclusivity in the digital ecosystem. Timeliness, accuracy, and coverage remain central for credibility [411]. By moving beyond English dominance, CLIR fosters collaboration and ensures high-quality information is accessible worldwide.

6 Challenges and Future Directions

6.1 CLIR Challenges

CLIR approaches are shaped by advances in multilingual NLP as well as monolingual IR. Challenges from both areas interact and create compound problems. For instance, we have seen that short queries in monolingual IR often contain ambiguity. In CLIR, additional uncertainty arises from translation or embedding, which can shift meaning and lead to the retrieval of irrelevant documents. Multilingual LLMs face further obstacles related to data, linguistic representation, model robustness, and generalisation. Below, we summarise the main challenges.

Cross-Linguality and Language. Short queries frequently lack context, making them ambiguous and reducing retrieval accuracy. Polysemy and homonymy are especially problematic across languages (as seen above in Figure 5). Without contextual cues, it is difficult to identify intended meanings, leading to semantic divergence. OOV terms also hinder retrieval, particularly proper nouns or newly-coined technical terms. Large training corpora partly address this, but low-resource languages remain unsupported. Dissimilar character sets add complications, since transliteration is inconsistent [25]. Morphological and syntactic variation across typologically distant languages makes alignment less reliable [323, 412]. Across languages, morphology is also relevant for the translation direction: translating from a language with a simple morphology to a more morphologically rich one tends to perform poorly compared to the other way around [25].

Language Representations. Languages differ not only in structure and vocabulary but also in semantic priorities. Some concepts are easily expressed in one language but rare or absent in another, especially across semantically distant pairs [323] (see Table 6, which illustrates translation challenges in Japanese and German). These differences hinder the creation of a language-agnostic embedding space. Shared embeddings reduce reliance on translation, but catastrophic forgetting, where a model loses performance on previously learned languages while adapting to new ones, remains a challenge [413]. Tokenisation also poses difficulties: inadequate segmentation increases OOV issues and fragments representations, especially in under-resourced languages [248].

Term	Translation Challenge Explanation
Japanese: わびさび (Wabi-sabi)	<p>Literal Components:</p> <ul style="list-style-type: none"> • <i>Wabi</i> (侘び) — rustic simplicity, quietness, subtle melancholy • <i>Sabi</i> (寂び) — the beauty of ageing, weathering, and impermanence <p>Overall Meaning: A worldview centred on the acceptance of transience and imperfection. It celebrates the beauty of things that are humble, weathered, and incomplete - something hard to express succinctly in English.</p>
German: Waldensamkeit	<p>Literal Components:</p> <ul style="list-style-type: none"> • <i>Wald</i> — forest • <i>Einsamkeit</i> — solitude, loneliness <p>Overall Meaning: A profound, peaceful feeling of solitude and connection with nature experienced while being alone in the forest, more emotional and poetic than a simple “lonely forest” or “solitude in woods”.</p>

Table 6: Examples of culturally embedded terms that defy literal translation.

Data and Resources. Data imbalance affects every stage of CLIR. High-resource languages dominate pre-training corpora (Figure 6 presents the imbalance of languages across online websites), skewing token representations and alignment. While high-quality parallel corpora are costly to build, machine-generated ones are cheaper but often distort meaning [365, 414]. Most datasets rely heavily on Wikipedia, enabling strong general-domain performance but weak results in specialised areas such as medicine or law. Dedicated resources for CLIR remain limited.

Bias. Multilingual models often perform well on high-resource languages but poorly on low-resource ones, leading to weak embeddings and alignment [182, 188, 415, 416]. Translation-based pipelines mitigate this through pivot languages, but embedding-based CLIR requires uniform representations across languages. Generative models also reflect demographic bias present in training data [417, 418]. Bias further affects evaluation, as skewed datasets or metric choices can distort results and give a misleading perception of system effectiveness [419–421].

Evaluation. Evaluation metrics for CLIR are largely adapted from monolingual IR and often fail to capture cross-lingual challenges. With the rise of LLMs, answer generation becomes central, yet metrics to evaluate it remain insufficient. Hallucinations where models produce inaccurate or nonsensical responses, pose particular risks when users rely on summaries rather than original documents. Trust declines further when fabricated references are included [171].

Engineering Challenges. CLIR involves resource-intensive steps such as translation, cross-lingual embedding, and retrieval. The high inference cost of multilingual LLMs exacerbate this, especially for interactive systems that require speed. Current pipelines combine monolingual and cross-lingual techniques, but subcomponents (e.g. query expansion and ranking) are often designed independently, complicating integration.

CLIR faces persistent challenges particularly due to the combination of applying monolingual information retrieval methods, which assume a uniform and balanced document corpus, to multilingual settings, where data, representations, and performance are often uneven, especially in low-resource languages. Addressing these interconnected issues is essential for progress.

6.2 Future Directions in CLIR

Cross-lingual information retrieval has seen significant progress in recent years with the rise of multilingual generative models and embedding-based retrieval techniques. This section covers some promising research directions:

- *Language-agnostic representations.* Advances in contrastive pre-training and multilingual embedding alignment are promising [180, 257], but models still underperform in low-resource and typologically diverse languages [323]. Future work should explore training objectives that enforce semantic consistency across languages at the sentence and document level.
- *Low-resource languages.* Addressing data imbalance remains essential [11]. Expanding training and pseudo-parallel corpora across diverse languages would enhance model performance. LLMs also show potential for generating training data, whether for queries, documents, or translation [414].
- *Multimodal CLIR.* Incorporating images, captions, and speech queries broadens CLIR applications. Expanding to cross-lingual multimodal inputs requires new pre-training strategies, benchmarks and evaluation metrics [422, 423].
- *Graph-based retrieval.* For complex knowledge-grounded CLIR, graph-based approaches are promising [424–426]. RAG frameworks could be extended to multilingual knowledge graphs for factual grounding and disambiguation.
- *Misinformation.* As CLIR systems increasingly incorporate generative models, risks of hallucination and inaccurate outputs remain a concern [66, 168]. Future research should focus on integrating robust fact-checking and hallucination-aware retrieval mechanisms [171], as well as developing models that can provide transparent references and assess the reliability of information sources across languages.
- *Disambiguation.* CLIR ambiguity arises from short queries and polysemy, which are especially problematic in multilingual contexts. Future systems could allow clarification of user intent through conversational search and build on query expansion methods [36, 37, 54].
- *Benchmarks and metrics.* Continued development of multilingual and multi-domain benchmarks is essential, building on datasets such as MIRACL [99]. Metrics are required to evaluate semantic drift, retrieval robustness, and QA performance.
- *Bias and representation robustness.* CLIR systems are susceptible to bias (e.g. stereotypical, cultural, linguistic) due to imbalances in training data, leading to disparities in retrieval performance. Recent work proposes methods such as OPTICAL (Optimal Transport Distillation) to transfer alignment from high-resource to low-resource languages [182] and adversarial training [188] to reduce embedding biases and improve zero-shot performance [427]. Another line of research addresses tokenisation challenges. Universal tokenisers trained across diverse languages have shown improved adaptability to unseen languages [428], while cross-lingual vocabulary transfer methods such as trans-tokenisation [429] initialise embeddings using semantically related tokens from high-resource languages. These strategies highlight the need for continued exploration of bias reduction and robust representation of learning across diverse linguistic contexts, especially beyond European languages.

Taken together, these directions highlight that although significant progress has been achieved in CLIR, substantial challenges remain. Addressing resource imbalance, bias, misinformation, and disambiguation, while extending systems to multimodal and graph-based contexts, will be key to advancing robust and equitable CLIR.

7 Conclusion

Cross-lingual information retrieval lies at the intersection of two rapidly growing fields: multilingual representation learning and information retrieval. As NLP technologies become increasingly integrated into everyday tools and services, the ability to access information across language boundaries has become more critical than ever. This survey has presented a comprehensive overview of CLIR systems, covering advances in representation learning, embedding alignment, retrieval techniques, evaluation methods, and system architectures. The discussion underscores that effective CLIR requires more than the adaptation of monolingual approaches or the direct reuse of multilingual models. Progress depends on innovations that address linguistic mismatch, resource imbalance, and the development of reliable evaluation frameworks. At the same time, CLIR plays a vital role in democratising access to knowledge. Since the majority of digital resources are concentrated in a small number of high-resource languages, equitable access relies on systems that are accurate, robust, and language-agnostic, with the potential to expand opportunities in science, education, healthcare, commerce, and many other domains.

Future development must confront persistent challenges. The scarcity of high-quality multilingual data, particularly for low-resource languages, remains a major bottleneck, making inclusive training corpora and stronger evaluation benchmarks essential. Sustained collaboration between academic and industrial communities will be crucial, along with attention to fairness, bias mitigation, and linguistic diversity in the system design. This survey provides both a foundation and a reference point for advancing the field. By uniting progress in multilingual modelling with innovations in information retrieval, CLIR has the capacity to transform global access to knowledge. Continued efforts will be necessary to ensure that these technologies serve all languages equitably, creating a future where language is no longer a barrier to information.

References

- [1] F. Wilfrid Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. Wiley, New York, 1979.
- [2] Karen Sparck Jones and Peter Willett, editors. *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1558604545.
- [3] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., USA, 1999. ISBN 020139829X.
- [4] BR Pevzner. Automatic translation of english text to the language of the pusto-nepusto-2 system. *Automatic Documentation and Mathematical Linguistics*, 3(4):40–48, 1969.
- [5] Gerard Salton. Experiments in multi-lingual information retrieval. Technical report, Cornell University, 1972.
- [6] Peter Schäuble and Páraic Sheridan. Cross-language information retrieval (clir) track overview. *NIST SPECIAL PUBLICATION SP*, pages 31–44, 1998.
- [7] CAROL PETERS. First results of the clef 2000 cross-language text retrieval system evaluation campaign. In *Working Notes for the CLEF 2000 Workshop. Lisboa*, 2000.
- [8] Language Magazine. Information inequality and the languages of the internet, 2015. URL <https://languagemagazine.com/2015/05/29/information-inequality-and-the-languages-of-the-internet/>. Accessed: 2025-07-03.
- [9] Andrew Trotman. Education equity and the digital divide. *ResearchGate*, 2000. URL https://www.researchgate.net/publication/228616386_Education_equity_and_the_digital_divide. Accessed: 2025-07-03.
- [10] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL <https://aclanthology.org/2020.acl-main.560/>.
- [11] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: A metric for quantifying llm performance across high and low-resource languages, 2024. URL <https://arxiv.org/abs/2404.11553>.

- [12] Tiya Vaj. Laser (language-agnostic sentence representations). Online article, 2023.
- [13] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, November 2021. doi: 10.18653/v1/2021.mrl-1.24. URL <https://aclanthology.org/2021.mrl-1.24>.
- [14] Chao-Wei Huang, Chen-An Li, Tsu-Yuan Hsu, Chen-Yu Hsu, and Yun-Nung Chen. Unsupervised multilingual dense retrieval via generative pseudo labeling. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 736–746, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-eacl.49/>.
- [15] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval, 2025. URL <https://arxiv.org/abs/2508.21038>.
- [16] Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. Zero-shot cross-lingual reranking with large language models for low-resource languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–656, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.59. URL <https://aclanthology.org/2024.acl-short.59/>.
- [17] Kailash A Hambarde and Hugo Proenca. Information retrieval: recent advances and beyond. *IEEE Access*, 11: 76581–76604, 2023.
- [18] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024. URL <https://arxiv.org/abs/2308.07107>.
- [19] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.*, 43(3), May 2025. ISSN 1046-8188. doi: 10.1145/3722552. URL <https://doi.org/10.1145/3722552>.
- [20] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, August 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL <http://dx.doi.org/10.1613/jair.1.11640>.
- [21] Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. Multilingual large language models: A systematic survey, 2024. URL <https://arxiv.org/abs/2411.11072>.
- [22] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jincheng Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025. URL <https://arxiv.org/abs/2405.10936>.
- [23] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11), April 2025. ISSN 2095-2236. doi: 10.1007/s11704-024-40579-4. URL <http://dx.doi.org/10.1007/s11704-024-40579-4>.
- [24] Jian-Yun Nie. *Cross-Language Information Retrieval*. Springer Cham, 2010.
- [25] Petra Galuščáková, Douglas W. Oard, and Suraj Nair. Cross-language information retrieval, 2022. URL <https://arxiv.org/abs/2111.05988>.
- [26] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, 2000. ISSN 0306-4573. doi: [https://doi.org/10.1016/S0306-4573\(99\)00056-4](https://doi.org/10.1016/S0306-4573(99)00056-4). URL <https://www.sciencedirect.com/science/article/pii/S0306457399000564>.
- [27] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001. doi: [https://doi.org/10.1002/1097-4571\(2000\)9999:9999<:AID-ASI1591>3.0.CO;2-R](https://doi.org/10.1002/1097-4571(2000)9999:9999<:AID-ASI1591>3.0.CO;2-R). URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/1097-4571%282000%299999%3A9999%3C%3A%3AAID-ASI1591%3E3.0.CO%3B2-R>.

- [28] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, November 1987. ISSN 0001-0782. doi: 10.1145/32206.32212. URL <https://doi.org/10.1145/32206.32212>.
- [29] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1), January 2012. ISSN 0360-0300. doi: 10.1145/2071389.2071390. URL <https://doi.org/10.1145/2071389.2071390>.
- [30] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792552. URL <https://doi.org/10.1145/792550.792552>.
- [31] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction, 2019. URL <https://arxiv.org/abs/1904.08375>.
- [32] Stephen E. Robertson. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- [33] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models, 2023. URL <https://arxiv.org/abs/2303.07678>.
- [34] Hiteshwar Kumar Azad and Akshay Deepak. Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5):1698–1735, 2019. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2019.05.009>. URL <https://www.sciencedirect.com/science/article/pii/S0306457318305466>.
- [35] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244, 12 1990. ISSN 0950-3846. doi: 10.1093/ijl/3.4.235. URL <https://doi.org/10.1093/ijl/3.4.235>.
- [36] J. J. Rocchio and G. Salton. Information search optimization and interactive retrieval techniques. In *Proceedings of the November 30–December 1, 1965, Fall Joint Computer Conference, Part I*, AFIPS ’65 (Fall, part I), page 293–305, New York, NY, USA, 1965. Association for Computing Machinery. ISBN 9781450378857. doi: 10.1145/1463891.1463926. URL <https://doi.org/10.1145/1463891.1463926>.
- [37] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [38] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. *SIGIR Forum*, 51(2): 168–175, August 2017. ISSN 0163-5840. doi: 10.1145/3130348.3130364. URL <https://doi.org/10.1145/3130348.3130364>.
- [39] M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3): 130–137, 1980.
- [40] Ming-Hung Hsu, Ming-Feng Tsai, and Hsin-Hsi Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. In Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan, and Donghong Ji, editors, *Information Retrieval Technology*, pages 1–13, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-46237-8.
- [41] Ming-Hung Hsu, Ming-Feng Tsai, and Hsin-Hsi Chen. Combining wordnet and conceptnet for automatic query expansion: A learning approach. In Hang Li, Ting Liu, Wei-Ying Ma, Tetsuya Sakai, Kam-Fai Wong, and Guodong Zhou, editors, *Information Retrieval Technology*, pages 213–224, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-68636-1.
- [42] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’01, page 120–127, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133316. doi: 10.1145/383952.383972. URL <https://doi.org/10.1145/383952.383972>.
- [43] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM ’01, page 403–410, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581134363. doi: 10.1145/502585.502654. URL <https://doi.org/10.1145/502585.502654>.

- [44] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990. doi: [https://doi.org/10.1002/\(SICI\)1097-4571\(199006\)41:4<288::AID-ASI8>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8>3.0.CO;2-H). URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199006%2941%3A4%3C288%3A%3AID-ASI8%3E3.0.CO%3B2-H>.
- [45] Anne Sihvonen and Pertti Vakkari. Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation*, 60(6):673–690, 2004. ISSN 0022-0418. doi: 10.1108/00220410410568151.
- [46] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 243–250, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581644. doi: 10.1145/1390334.1390377. URL <https://doi.org/10.1145/1390334.1390377>.
- [47] Claudio Carpineto, Giovanni Romano, and Vittorio Giannini. Improving retrieval feedback with multiple term-ranking function combination. *ACM Trans. Inf. Syst.*, 20(3):259–290, July 2002. ISSN 1046-8188. doi: 10.1145/568727.568728. URL <https://doi.org/10.1145/568727.568728>.
- [48] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582416. URL <https://doi.org/10.1145/582415.582416>.
- [49] Youjin Chang, Iadh Ounis, and Minkoo Kim. Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management*, 42(2):453–468, 2006. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2005.03.025>. URL <https://www.sciencedirect.com/science/article/pii/S0306457305000567>.
- [50] Andrea Bernardini and Claudio Carpineto. FUB at TREC 2008 relevance feedback track: Extending rocchio with distributional term analysis. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008*, volume 500-277 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2008. URL <http://trec.nist.gov/pubs/trec17/papers/fondazione.rf.rev.pdf>.
- [51] W.S. Wong, R.W.P. Luk, H.V. Leong, K.S. Ho, and D.L. Lee. Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing & Management*, 44(3):1086–1116, 2008. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2007.12.002>. URL <https://www.sciencedirect.com/science/article/pii/S0306457307002191>.
- [52] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using smart: Trec 3. In *Text Retrieval Conference*, 1994. URL <https://api.semanticscholar.org/CorpusID:14683127>.
- [53] Alexander M. Robertson and Peter Willett. A comparison of spelling-correction methods for the identification of word forms in historical text databases*. *Literary and Linguistic Computing*, 8(3):143–152, 01 1993. ISSN 0268-1145. doi: 10.1093/lilc/8.3.143. URL <https://doi.org/10.1093/lilc/8.3.143>.
- [54] E.N. Efthimiadis. Query expansion. In M.E. Williams, editor, *Annual review of information science and technology*, pages 121–187. Information Today, Medford, NJ, 1996.
- [55] G. Salton and M. E. Lesk. The smart automatic document retrieval systems—an illustration. *Commun. ACM*, 8(6):391–398, June 1965. ISSN 0001-0782. doi: 10.1145/364955.364990. URL <https://doi.org/10.1145/364955.364990>.
- [56] Gerard Salton. The evaluation of automatic retrieval procedures—selected test results using the smart system. *American Documentation*, 16(3):209–222, 1965.
- [57] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.
- [58] Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. Umass at trec 2004: Novelty and hard. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC-13)*, 2004.

- [59] Minghan Li, Xinxuan Lv, Junjie Zou, Tongna Chen, Chao Zhang, Suchao An, Ercong Nie, and Guodong Zhou. Query expansion in the age of pre-trained and large language models: A comprehensive survey, 2025. URL <https://arxiv.org/abs/2509.07794>.
- [60] Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. Large language model based long-tail query rewriting in taobao search, 2024. URL <https://arxiv.org/abs/2311.03758>.
- [61] Minghan Li, Honglei Zhuang, Kai Hui, Zhen Qin, Jimmy Lin, Rolf Jagerman, Xuanhui Wang, and Michael Bendersky. Can query expansion improve generalization of strong cross-encoder rankers? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2321–2326, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657979. URL <https://doi.org/10.1145/3626772.3657979>.
- [62] Vincent Claveau. Neural text generation for query expansion in information retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 202–209, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391153. doi: 10.1145/3486622.3493957. URL <https://doi.org/10.1145/3486622.3493957>.
- [63] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 7–16, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2063584. URL <https://doi.org/10.1145/2063576.2063584>.
- [64] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.99. URL <https://aclanthology.org/2023.acl-long.99/>.
- [65] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models, 2023. URL <https://arxiv.org/abs/2305.03653>.
- [66] Kenya Abe, Kunihiro Takeoka, Makoto P. Kato, and Masafumi Oyamada. Llm-based query expansion fails for unfamiliar and ambiguous queries, 2025. URL <https://arxiv.org/abs/2505.12694>.
- [67] Wenjing Zhang, Zhaoxiang Liu, Kai Wang, and Shiguo Lian. Query expansion and verification with large language model for information retrieval. In De-Shuang Huang, Zhanjun Si, and Chuanlei Zhang, editors, *Advanced Intelligent Computing Technology and Applications*, pages 341–351, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-97-5672-8.
- [68] Shadi Saleh and Pavel Pecina. Term selection for query expansion in medical cross-lingual information retrieval. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, pages 507–522, Cham, 2019. Springer International Publishing. ISBN 978-3-030-15712-8.
- [69] Benoît Gaillard, Jean-Léon Bouraoui, Emilie Guimier de Neef, and Malek Boualem. Query expansion for cross language information retrieval improvement. In *2010 Fourth International Conference on Research Challenges in Information Science (RCIS)*, pages 337–342, 2010. doi: 10.1109/RCIS.2010.5507393.
- [70] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, page 84–91, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918363. doi: 10.1145/258525.258540. URL <https://doi.org/10.1145/258525.258540>.
- [71] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 64–71, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.290958. URL <https://doi.org/10.1145/290941.290958>.

- [72] Paul McNamee and James Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 159–166, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135610. doi: 10.1145/564376.564406. URL <https://doi.org/10.1145/564376.564406>.
- [73] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.
- [74] Toshitaka Kuwa, Shigehiko Schamoni, and Stefan Riezler. Embedding meta-textual information for improved learning to rank. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 5558–5568, 2020.
- [75] Tsegaye Misikir Tashu, Eduard-Raul Kontos, Matthia Sabatelli, and Matias Valdenegro-Toro. Mapping transformer leveraged embeddings for cross-lingual document representation. *arXiv preprint arXiv:2401.06583*, 2024.
- [76] Xiyang Hu, Xinchu Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. Language agnostic multilingual information retrieval with contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9133–9146, 2023. doi: 10.18653/v1/2023.findings-acl.581. URL <https://aclanthology.org/2023.findings-acl.581/>.
- [77] Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. mDAPT: Multilingual domain adaptive pretraining in a single model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418, 2021. doi: 10.18653/v1/2021.findings-emnlp.290. URL <https://aclanthology.org/2021.findings-emnlp.290/>.
- [78] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Learning domain-specialised representations for cross-lingual biomedical entity linking, 2021. URL <https://arxiv.org/abs/2105.14398>.
- [79] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained on document-aligned comparable corpora. *Information Retrieval Journal*, 16(3): 331–368, 2013. doi: 10.1007/s10791-012-9200-5.
- [80] David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of EMNLP 2009*, pages 880–889, 2009.
- [81] Lisa Posch, Arnim Bleier, Philipp Schaer, and Markus Strohmaier. The polylingual labeled topic model. In *Proceedings of COLING 2014*, pages 2982–2992, 2015.
- [82] Siddhartha Devapujula, Sagar Arora, and Sumit Borar. Learning to rank broad and narrow queries in e-commerce. *arXiv preprint arXiv:1907.01549*, 2019. URL <https://arxiv.org/abs/1907.01549>.
- [83] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. On application of learning to rank for e-commerce search. In *Proceedings of SIGIR 2017 e-Commerce Workshop*, 2017. URL <https://arxiv.org/abs/1903.04263>.
- [84] Bryan Zhang, Liling Tan, and Amita Misra. Evaluating machine translation in cross-lingual e-commerce search. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, pages 322–334, Orlando, USA, 2022. AMTA.
- [85] William S Cooper, Fredric C Gey, and Daniel P Dabney. Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 198–210, 1992.
- [86] Norbert Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems (TOIS)*, 7(3):183–204, 1989.
- [87] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.
- [88] Chris J. C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report msr-tr-2010-82, Microsoft Research, 2010. URL http://research.microsoft.com/en-us/um/people/cburges/tech_reports/MSR-TR-2010-82.pdf.

- [89] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 129–136, 2007. doi: 10.1145/1273496.1273513.
- [90] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1192–1199, 2008. doi: 10.1145/1390156.1390306.
- [91] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
- [92] D. Manning, Christopher Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [93] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3): 225–331, 2009. doi: 10.1561/15000000016.
- [94] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, 2020.
- [95] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021. URL <https://arxiv.org/abs/2010.11934>.
- [96] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL <https://arxiv.org/abs/2002.10957>.
- [97] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset, 2022. URL <https://arxiv.org/abs/2108.13897>.
- [98] Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. Xor qa: Cross-lingual open-retrieval question answering, 2021. URL <https://arxiv.org/abs/2010.11856>.
- [99] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tac1_a_00595. URL <https://aclanthology.org/2023.tac1-1.63/>.
- [100] Eugene Yang, Dawn J. Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Advances in Information Retrieval: 46th European Conference on IR (ECIR)*, pages 50–65, 2024. doi: 10.1007/978-3-031-56060-6_4.
- [101] Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. Neuralmind-unicamp at 2022 trec neuclir: Large boring rerankers for cross-lingual retrieval, 2023. URL <https://arxiv.org/abs/2303.16145>.
- [102] Eugene Yang, Dawn Lawrie, and James Mayfield. Hltcoe at trec 2023 neuclir track. *arXiv preprint arXiv:2404.08118*, 2024.
- [103] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 2387–2392, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531863.
- [104] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL <https://aclanthology.org/2023.emnlp-main.923/>.

- [105] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*, 2023.
- [106] Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler, and Kai Hui. Parade: Passage ranking using demonstrations with large language models, 2023. URL <https://arxiv.org/abs/2310.14408>.
- [107] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL <https://arxiv.org/abs/2210.11416>.
- [108] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. URL <https://arxiv.org/abs/2310.16944>.
- [109] OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022. Blog post announcing ChatGPT by OpenAI.
- [110] Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. Attention in large language models yields efficient zero-shot re-rankers. In *ICLR 2025 Workshop*, 2025.
- [111] Herv   D  jean, St  phane Clinchant, and Thibault Formal. A thorough comparison of cross-encoders and llms for reranking splade, 2024. URL <https://arxiv.org/abs/2403.10407>.
- [112] Anonymous. Multilingual open qa on the mia shared task. *arXiv preprint arXiv:2501.04153*, 2025.
- [113] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL <https://aclanthology.org/2020.acl-main.653/>.
- [114] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. On application of learning to rank for e-commerce search, 2019.
- [115] Gordon V. Cormack, Charles L. A. Clarke, and Stefan B  ttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR ’09*, 2009.
- [116] Rrf ranker — zilliz cloud developer hub. <https://docs.zilliz.com/docs/reranking-rrf>, 2025. Describes use of Reciprocal Rank Fusion for multilingual and hybrid search.
- [117] Hybrid search scoring via reciprocal rank fusion — azure ai search. <https://learn.microsoft.com/en-us/azure/search/hybrid-search-ranking>, 2025. Explains RRF for combining vector and lexical search results.
- [118] Laura Dietz, Hannah Bast, Shubham Chatterjee, Jeff Dalton, Edgar Meij, and Arjen de Vries. Ecir 23 tutorial: Neuro-symbolic approaches for information retrieval. In *European Conference on Information Retrieval*, pages 324–330. Springer, 2023.
- [119] Le Zhang, Damianos Karakos, William Hartmann, Manaj Srivastava, Lee Tarlin, David Akodes, Sanjay Krishna Gouda, Numra Bathool, Lingjun Zhao, Zhuolin Jiang, Richard Schwartz, and John Makhoul. The 2019 BBN cross-lingual information retrieval system. In Kathy McKeown, Douglas W. Oard, Elizabeth, and Richard Schwartz, editors, *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 44–51, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-55-9. URL <https://aclanthology.org/2020.clssts-1.8/>.
- [120] Nick Craswell, Onno Zoeter, and Michael Taylor. An experimental comparison of click position-bias models. In *WSDM ’08*, pages 87–94, 2008.

- [121] Olivier Chapelle and Yi Zhang. A dynamic bayesian network click model for web search ranking. In *WWW '09*, pages 1–10, 2009.
- [122] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 610–618, 2018.
- [123] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7–es, April 2007. ISSN 1046-8188. doi: 10.1145/1229179.1229181. URL <https://doi.org/10.1145/1229179.1229181>.
- [124] Md Aminul Islam and Ahmed Sayeed Faruk. Prompt-based llms for position bias-aware reranking in personalized recommendations. In *arXiv preprint arXiv:2505.04948*, 2025.
- [125] Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *NAACL 2024 Long Papers*, 2024. doi: 10.18653/v1/2024.naacl-long.129.
- [126] Jade Goldstein and Jaime Carbonell. Using mmr, diversity-based reranking for reordering query results. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998. doi: 10.1145/290941.291025.
- [127] Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronimo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and Xinyu Zhang. Simple yet effective neural ranking and reranking baselines for cross-lingual information retrieval, 2023. URL <https://arxiv.org/abs/2304.01019>.
- [128] Lidan Wang, Jimmy J. Lin, and Donald Metzler. A cascade ranking model for efficient ranked retrieval. In *SIGIR '11*, pages 105–114, 2011.
- [129] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In *SIGIR '17*, pages 445–454, 2017.
- [130] Robert Honig, Jan Ackermann, and Mingyuan Chi. Bi-encoder cascades for efficient image search. In *ICCV Workshop RCV '23*, 2023.
- [131] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Ranked list truncation for large language model-based re-ranking. In *SIGIR '24*, 2024.
- [132] Jianghong Zhou and Eugene Agichtein. Rlirank: Learning to rank with reinforcement learning for dynamic search. *arXiv preprint arXiv:2105.10124*, 2021.
- [133] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [134] Tim Baumgärtner, Leonardo F. R. Ribeiro, Nils Reimers, and Iryna Gurevych. Incorporating relevance feedback for information-seeking retrieval using few-shot document re-ranking. In *EMNLP Findings 2022*, 2022.
- [135] Minh Nguyen, Toan Quoc Nguyen, Kishan KC, Zeyu Zhang, and Thuy Vu. Reinforcement learning from answer reranking feedback for retrieval-augmented answer generation. In *Proceedings of INTERSPEECH*, 2024.
- [136] Grace Hui Yang and Ian Soboroff. Trec 2016 dynamic domain track overview. In *TREC*, 2016.
- [137] Jurek Leonhardt, Koustav Rudra, and Avishek Anand. Extractive explanations for interpretable text ranking. *ACM Transactions on Information Systems*, 41(4):1–27, 2023.
- [138] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL-HLT '19*, pages 3543–3556, 2019.
- [139] Sarah Wiegrefe and Yuval Pinter. Attention is not not explanation. In *NAACL-HLT '19*, pages 854–863, 2019.
- [140] Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. Seat: Stable and explainable attention, 2022. URL <https://arxiv.org/abs/2211.13290>.

- [141] Saran Pandian, Debasis Ganguly, and Sean MacAvaney. Evaluating the explainability of neural rankers, 2024. URL <https://arxiv.org/abs/2403.01981>.
- [142] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey, 2022. URL <https://arxiv.org/abs/2211.02405>.
- [143] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, 2020.
- [144] Keshav Santhanam, Omar Khattab, Theodoros Rekatsinas, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *NAACL*, 2022.
- [145] Yee Seng Chan and Hwee Tou Ng. MAXSIM: A maximum similarity metric for machine translation evaluation. In Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL <https://aclanthology.org/P08-1007/>.
- [146] Zhuyun Dai and Jamie Callan. Context-aware term weighting for first-stage passage retrieval. In *SIGIR*, pages 985–988, 2019. doi: 10.1145/3397271.3401204.
- [147] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *SIGIR*, 2021. arXiv preprint [arXiv:2107.05720](https://arxiv.org/abs/2107.05720).
- [148] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. In *SIGIR*, 2021. arXiv preprint [arXiv:2109.10086](https://arxiv.org/abs/2109.10086).
- [149] Eunseong Choi, Sunkyung Lee, Minjin Choi, Hyeseon Ko, Young-In Song, and Jongwuk Lee. Spade: Improving sparse representations using a dual document encoder for first-stage retrieval. In *arXiv preprint arXiv:2209.05917*, 2022.
- [150] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gattford. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST, 1995.
- [151] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas Oard. Learning a sparse representation model for neural clir. In *Proceedings of DESIRES 2022: Design of Experimental Search & Information Retrieval Systems*, 2022.
- [152] Hiroki Iida and Naoaki Okazaki. Unsupervised domain adaptation for sparse retrieval by filling vocabulary and word frequency gaps. *arXiv preprint arXiv:2211.03988*, 2022.
- [153] Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):1–35, 2023.
- [154] Weaviate Team. Hybrid search explained, 2025. Weaviate blog.
- [155] Yiqun Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. In *TACL*, 2020.
- [156] Luyu Gao, Zhuyun Dai, and Jamie Callan. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *NAACL*, 2021.
- [157] Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and Philip Yu. Dense hierarchical retrieval for open-domain question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 188–200, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.19. URL <https://aclanthology.org/2021.findings-emnlp.19/>.
- [158] Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. An efficient memory-augmented transformer for knowledge-intensive nlp tasks. In *EMNLP*, 2022.
- [159] Suyu Ge, Chenyan Xiong, Corby Rosset, Arnold Overwijk, Jiawei Han, and Paul Bennett. Augmenting zero-shot dense retrievers with plug-in mixture-of-memories (moma). *arXiv preprint arXiv:2302.03754*, 2023.

- [160] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL <https://aclanthology.org/P17-1171/>.
- [161] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfay, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL <https://aclanthology.org/2021.eacl-main.74/>.
- [162] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [163] Ryu Iida, Canasai Kruengkrai, Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloezer. Exploiting background knowledge in compact answer generation for why-questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):142–151, Jul. 2019. doi: 10.1609/aaai.v33i01.3301142. URL <https://ojs.aaai.org/index.php/AAAI/article/view/3779>.
- [164] Travis Goodwin, Max Savary, and Dina Demner-Fushman. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3215–3226, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.289. URL <https://aclanthology.org/2020.findings-emnlp.289/>.
- [165] Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. Joint learning of answer selection and answer summary generation in community question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7651–7658, Apr. 2020. doi: 10.1609/aaai.v34i05.6266. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6266>.
- [166] Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. Answer generation for retrieval-based question answering systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4276–4282, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.374. URL <https://aclanthology.org/2021.findings-acl.374/>.
- [167] Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. Cross-lingual open-domain question answering with answer sentence generation, 2022. URL <https://arxiv.org/abs/2110.07150>.
- [168] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, chapter Question Answering, Information Retrieval, and Retrieval-Augmented Generation. Forthcoming, 3rd edition, 2025.
- [169] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 06 2024. ISSN 2161-7201. doi: 10.1093/jla/laae003. URL <https://doi.org/10.1093/jla/laae003>.
- [170] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models’ reluctance to express uncertainty, 2024. URL <https://arxiv.org/abs/2401.06730>.
- [171] Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models, 2023. URL <https://arxiv.org/abs/2302.05578>.
- [172] Chia-Hsuan Lee and Hung-Yi Lee. Cross-lingual transfer learning for question answering, 2019. URL <https://arxiv.org/abs/1907.06042>.

- [173] Carlos Lassance. Extending english ir methods to multi-lingual ir. *arXiv preprint arXiv:2302.14723*, 2023.
- [174] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *ACL*, 2022.
- [175] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4513–4525. ACL, 2020.
- [176] Yi Zhang, Payal Bajaj, Xiao Ma, Akari Asai, Chenyan Xiong, Jamie Callan, and Hannaneh Hajishirzi. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.
- [177] Zhucheng Tu and Sarguna Janani Padmanabhan. 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)*, pages 100–107, Seattle, USA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.mia-1.10. URL <https://aclanthology.org/2022.mia-1.10/>.
- [178] Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.597. URL <https://aclanthology.org/2022.emnlp-main.597/>.
- [179] Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*, 2025. Published April 2025.
- [180] Luyu Gao, Haoyu Zhong, Jamie Callan, Longqing Xiong, and Wen-tau Yih. Cocondenser: Contrastive pretraining with hard negatives for dense passage retrieval. In *EMNLP*, 2022.
- [181] Shuo Sun and Kevin Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.340. URL <https://aclanthology.org/2020.emnlp-main.340/>.
- [182] Zhiqi Huang, Puxuan Yu, and James Allan. 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)*, pages 1048–1056, 2023. doi: 10.1145/3539597.3570468.
- [183] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL <https://arxiv.org/abs/2112.09118>.
- [184] Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval, 2024. URL <https://arxiv.org/abs/2311.05800>.
- [185] Hongliang Fei, Tan Yu, and Ping Li. Cross-lingual cross-modal pretraining for multimodal retrieval. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.285. URL <https://aclanthology.org/2021.naacl-main.285/>.
- [186] Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. Multimodal llm enhanced cross-lingual cross-modal retrieval, 2024. URL <https://arxiv.org/abs/2409.19961>.
- [187] Layne Berry, Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Hung yi Lee, and David Harwath. M-speechclip: Leveraging large-scale, pre-trained models for multilingual speech to image retrieval, 2023. URL <https://arxiv.org/abs/2211.01180>.

- [188] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. Unsupervised cross-lingual information retrieval using monolingual data only. *arXiv preprint arXiv:1805.00879*, May 2018.
- [189] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25:149–183, 2022.
- [190] James Gore, Joseph Polletta, and Behrooz Mansouri. Crossmath: Towards cross-lingual math information retrieval. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 101–105, 2024.
- [191] Eugene Yang, Dawn Lawrie, and James Mayfield. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2368–2373. ACM, July 2024. doi: 10.1145/3626772.3657955. URL <http://dx.doi.org/10.1145/3626772.3657955>.
- [192] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, 45(1), 2012. ISSN 0360-0300. doi: 10.1145/2379776.2379777. URL <https://doi.org/10.1145/2379776.2379777>.
- [193] Christopher Fox. A stop list for general text. *SIGIR Forum*, 24(1–2):19–21, September 1989. ISSN 0163-5840. doi: 10.1145/378881.378888. URL <https://doi.org/10.1145/378881.378888>.
- [194] Ferhan Ture and Jimmy Lin. Exploiting representations from statistical machine translation for cross-language information retrieval. *ACM Trans. Inf. Syst.*, 32(4), October 2014. ISSN 1046-8188. doi: 10.1145/2644807. URL <https://doi.org/10.1145/2644807>.
- [195] David A. Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, page 49–57, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917928. doi: 10.1145/243199.243212. URL <https://doi.org/10.1145/243199.243212>.
- [196] Lisa Ballesteros and W. Bruce Croft. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications*, DEXA '96, page 791–801, Berlin, Heidelberg, 1996. Springer-Verlag. ISBN 354061656X.
- [197] David A. Hull. Using structured queries for disambiguation in cross-language information retrieval. In *Proceedings of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, pages 84–98, Stanford, CA, 1997. AAAI Press.
- [198] Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 55–63, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.290957. URL <https://doi.org/10.1145/290941.290957>.
- [199] Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, page 338–344, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136463. doi: 10.1145/860435.860497. URL <https://doi.org/10.1145/860435.860497>.
- [200] Gina-Anne Levow and Douglas W. Oard. Translingual topic tracking with prise. In *Working Notes of the 3rd Topic Detection and Tracking Workshop (TDT-3)*. National Institutes of Standards and Technology, 2000.
- [201] Tim Leek, Hubert Jin, Sreenivasa Sista, and Richard Schwartz. The bbn cross-lingual topic detection and tracking system. In *Working Notes of the 3rd Topic Detection and Tracking Workshop (TDT-3)*. National Institutes of Standards and Technology, 2000.
- [202] Jinxi Xu and Ralph Weischedel. Empirical studies on the impact of lexical resources on clir performance. *Information Processing and Management*, 41(3):475–487, May 2005. ISSN 0306-4573. doi: 10.1016/j.ipm.2004.06.009. URL <https://doi.org/10.1016/j.ipm.2004.06.009>.
- [203] Salim Roukos, David Graff, and Dan Melamed. Hansard french/english (ldc95t20). Linguistic Data Consortium, Web Download, 1995. URL <https://catalog.ldc.upenn.edu/LDC95T20>. LDC95T20.

- [204] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005. URL <https://aclanthology.org/2005.mtsummit-papers.11/>.
- [205] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Poulliquen. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL <https://aclanthology.org/L16-1561/>.
- [206] Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. Evaluation of the Bible as a resource for cross-language information retrieval. In Andreas Witt, Gilles Sérasset, Susan Armstrong, Jim Breen, Ulrich Heid, and Felix Sasaki, editors, *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74, Sydney, Australia, July 2006. Association for Computational Linguistics. URL <https://aclanthology.org/W06-1009/>.
- [207] Martin Braschler and Peter Schäuble. Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval*, 3(3):273–284, October 2000. ISSN 1386-4564. doi: 10.1023/A:1026525127581. URL <https://doi.org/10.1023/A:1026525127581>.
- [208] Isabelle Moulinier and Hugo Molina-Salgado. Thomson legal and regulatory experiments for clef 2002. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval*, pages 155–163, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45237-9.
- [209] Martin Franz, J. Mccarley, and Salim Roukos. Ad hoc and multilingual information retrieval at ibm. In *Proceedings of trec Conference*, 1999.
- [210] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990. ISSN 0891-2017.
- [211] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. ISSN 0891-2017.
- [212] Felix Stahlberg. Neural machine translation: A review and survey, 2020. URL <https://arxiv.org/abs/1912.02047>.
- [213] Shereen A. Mohamed, Ashraf A. Elsayed, Y. F. Hassan, and Mohamed A. Abdou. Neural machine translation: past, present, and future. *Neural Comput. Appl.*, 33(23):15919–15931, December 2021. ISSN 0941-0643. doi: 10.1007/s00521-021-06268-0. URL <https://doi.org/10.1007/s00521-021-06268-0>.
- [214] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11), February 2023. ISSN 0360-0300. doi: 10.1145/3567592. URL <https://doi.org/10.1145/3567592>.
- [215] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5), September 2020. ISSN 0360-0300. doi: 10.1145/3406095. URL <https://doi.org/10.1145/3406095>.
- [216] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper.pdf.
- [217] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. Dissecting contextual word embeddings: Architecture and representation, 2018. URL <https://arxiv.org/abs/1808.08949>.
- [218] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, October 2013. Association for Computational Linguistics.

- [219] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi, editors, *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL <https://aclanthology.org/W14-4012/>.
- [220] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*, 2015. URL <http://arxiv.org/abs/1409.0473>.
- [221] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [222] Philip Gage. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February 1994. ISSN 0898-9788.
- [223] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162/>.
- [224] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. Bpe-dropout: Simple and effective subword regularization, 2020. URL <https://arxiv.org/abs/1910.13267>.
- [225] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012. doi: 10.1109/ICASSP.2012.6289079.
- [226] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation, 2016. URL <https://arxiv.org/abs/1609.08144>.
- [227] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL <https://aclanthology.org/D18-2012/>.
- [228] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. Transfer learning approaches for building cross-language dense retrieval models, 2022. URL <https://arxiv.org/abs/2201.08471>.
- [229] Eugene Yang, Dawn J. Lawrie, Paul McNamee, and James Mayfield. Extending translate-train for colbert-x to african language clir, 2024. URL <https://arxiv.org/abs/2404.08134>.
- [230] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.12. URL <https://aclanthology.org/2021.mrl-1.12/>.
- [231] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL <https://aclanthology.org/W17-3204/>.

- [232] Djoerd Hiemstra and Wessel Kraaij. Twenty-one at trec-7: ad-hoc and cross-language track. In E.M Voorhees and D.K. Harman, editors, *Proceedings of the seventh Text Retrieval Conference (TREC)*, NIST Special Publications, pages 227–238, United States, 1999. National Institute of Standards and Technology. Seventh Text REtrieval Conference, TREC-7 1998 ; Conference date: 09-11-1999 Through 11-11-1999.
- [233] Kazuaki Kishida and Noriko Kando. Two-stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: An experiment at clef 2003. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 253–262. Springer, 2003.
- [234] Noriko Kando, Kuang-hua Chen, and Kazuaki Kishida. Two-stage refinement of transitive query translation with english disambiguation for cross-language information retrieval: An experiment at clef 2004. (*No Title*), 2005.
- [235] Kazuaki Kishida and Noriko Kando. A hybrid approach to query and document translation using a pivot language for cross-language information retrieval. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories*, pages 93–101, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-45700-8.
- [236] Wessel Kraaij and Franciska MG de Jong. Transitive probabilistic clir models. In *7th International Conference on Computer-Assisted Information Retrieval, RIAO 2004:(Recherche d’Information et ses Applications)*, pages 69–81. Centre de Hautes Etudes Internationales d’Informatique Documentaire (CID), 2004.
- [237] Kazuaki Kishida. Technical issues of cross-language information retrieval: a review. *Information processing & management*, 41(3):433–455, 2005.
- [238] Aitao Chen and Fredric C. Gey. Experiments on cross-language and patent retrieval at ntcir-3 workshop. In *Proceedings of the Third NTCIR Workshop*, 2003.
- [239] Wen-Cheng Lin and Hsin-Hsi Chen. Description of ntu approach to ntcir3 multilingual information retrieval. In *Proceedings of the Third NTCIR Workshop*, 2003.
- [240] Fredric C. Gey, Hailing Jiang, Aitao Chen, and Ray R. Larson. Manual queries and machine translation in cross-language retrieval and interactive retrieval with cheshire ii at trec-7. In *Proceedings of the seventh Text Retrieval Conference (TREC)*, TREC’98, pages 463–476, 1998.
- [241] Tim Gollins and Mark Sanderson. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’01*, page 90–95, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133316. doi: 10.1145/383952.383965. URL <https://doi.org/10.1145/383952.383965>.
- [242] Raija Lehtokangas, Eija Airio, and Kalervo Järvelin. Transitive dictionary translation challenges direct dictionary translation in clir. *Information Processing & Management*, 40(6):973–988, 2004. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2003.10.005>. URL <https://www.sciencedirect.com/science/article/pii/S0306457303000864>.
- [243] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9). URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>.
- [244] Michael L. Littman, Susan T. Dumais, and Thomas K. Landauer. *Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing*, pages 51–62. Springer US, Boston, MA, 1998. ISBN 978-1-4615-5661-9. doi: 10.1007/978-1-4615-5661-9_5. URL https://doi.org/10.1007/978-1-4615-5661-9_5.
- [245] Michael W. Berry and Paul G. Young. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429, 1995. ISSN 00104817. URL <http://www.jstor.org/stable/30200366>.
- [246] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, page 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
- [247] Anja Reusch and Yonatan Belinkov. Reverse-engineering the retrieval process in genir models, 2025. URL <https://arxiv.org/abs/2503.19715>.

- [248] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. A survey of multilingual large language models. *Patterns*, 6(1):101118, 2025. ISSN 2666-3899. doi: <https://doi.org/10.1016/j.patter.2024.101118>. URL <https://www.sciencedirect.com/science/article/pii/S2666389924002903>.
- [249] Daniil Gurgurov, Tanja Bäuml, and Tatiana Anikina. Multilingual large language models and curse of multilinguality, 2024. URL <https://arxiv.org/abs/2406.10602>.
- [250] Sumanth Doddapaneni, Gowtham Ramesh, Mitesh Khapra, Anoop Kunchukuttan, and Pratyush Kumar. A primer on pretrained multilingual language models. *ACM Comput. Surv.*, 57(9), May 2025. ISSN 0360-0300. doi: 10.1145/3727339. URL <https://doi.org/10.1145/3727339>.
- [251] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- [252] Qwen Team. Qwen3-embedding (including qwen3-reranker-8b). <https://qwenlm.github.io/blog/qwen3/>, May 2025. URL <https://qwenlm.github.io/blog/qwen3/>. Includes Qwen3-Reranker-8B reranker model.
- [253] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024. URL <https://arxiv.org/abs/2402.05672>.
- [254] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.813. URL <https://aclanthology.org/2023.emnlp-main.813/>.
- [255] Zach Nussbaum and Brandon Duderstadt. Training sparse mixture of experts text embedding models, 2025. URL <https://arxiv.org/abs/2502.07972>.
- [256] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025. URL <https://arxiv.org/abs/2503.07891>.
- [257] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. 2020.
- [258] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020. URL <https://arxiv.org/abs/2001.08210>.
- [259] 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, Sema 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. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.
- [260] David Uthus, Santiago Ontañón, Joshua Ainslie, and Mandy Guo. mlongt5: A multilingual and efficient text-to-text transformer for longer sequences, 2023. URL <https://arxiv.org/abs/2305.11129>.

- [261] Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model, 2024. URL <https://arxiv.org/abs/2402.07827>.
- [262] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2022. URL <https://arxiv.org/abs/2112.10668>.
- [263] Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. URL <https://ai.meta.com/blog/meta-llama-3/>.
- [264] Mistral AI. Model card for mistral-large-instruct-2407, 2024. URL <https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>.
- [265] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klammer, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Froberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névél, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes,

- Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyeade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel León Perrián, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sängner, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljeic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaronsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>.
- [266] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL <https://arxiv.org/abs/2403.04652>.
- [267] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gierler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madeline

- Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubei, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- [268] Google DeepMind. Gemini 2.5 pro model card. Technical report, 2025. URL <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf>.
- [269] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, 2024. URL <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf>.
- [270] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. URL <https://arxiv.org/abs/2305.10403>.
- [271] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaoqun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiusi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe

- Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanxia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>.
- [272] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL <https://arxiv.org/abs/2403.08295>.
- [273] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. Polym: An open source polyglot large language model, 2023. URL <https://arxiv.org/abs/2307.06018>.
- [274] Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Dan Su, Chen Zhu, Deepak Narayanan, Aastha Jhunjhunwala, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ameya Mahabaleshwarkar, Osvald Nitski, Annika Brundyn, James Maki, Miguel Martinez, Jiaxuan You, John Kamalu, Patrick LeGresley, Denys Fridman, Jared Casper, Ashwath Aithal, Oleksii Kuchaiev, Mohammad Shoeybi, Jonathan Cohen, and Bryan Catanzaro. Nemotron-4 15b technical report, 2024. URL <https://arxiv.org/abs/2402.16819>.
- [275] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: [https://doi.org/10.1016/S0079-7421\(08\)60536-8](https://doi.org/10.1016/S0079-7421(08)60536-8). URL <https://www.sciencedirect.com/science/article/pii/S0079742108605368>.
- [276] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyei Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey, 2024. URL <https://arxiv.org/abs/2404.16789>.
- [277] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: [10.1073/pnas.1611835114](https://doi.org/10.1073/pnas.1611835114). URL <https://www.pnas.org/doi/abs/10.1073/pnas.1611835114>.

- [278] 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. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- [279] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [280] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- [281] Marcus Williams, Micah Carroll, Adhyayan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback, 2025. URL <https://arxiv.org/abs/2411.02306>.
- [282] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french, 2020. URL <https://arxiv.org/abs/1912.05372>.
- [283] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model, 2019. URL <https://arxiv.org/abs/1912.09582>.
- [284] Sello Ralethe. Adaptation of deep bidirectional transformers for Afrikaans language. In Nicoletta Calzolari, Frédéric B  chet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H  l  ne Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2475–2478, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.301/>.
- [285] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- [286] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL <https://aclanthology.org/2023.acl-long.693/>.
- [287] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL <https://aclanthology.org/2021.mrl-1.11/>.

- [288] Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. Evaluation of the bible as a resource for cross-language information retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, MLRI '06, page 68–74, USA, 2006. Association for Computational Linguistics. ISBN 1932432825.
- [289] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- [290] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, March 2003. ISSN 1532-4435.
- [291] Alexis Conneau and Guillaume Lample. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [292] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora, 2021. URL <https://arxiv.org/abs/2012.15674>.
- [293] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1252. URL <https://aclanthology.org/D19-1252/>.
- [294] Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9386–9393, Apr. 2020. doi: 10.1609/aaai.v34i05.6480. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6480>.
- [295] W. L. Taylor. "cloze procedure": a new tool for measuring readability. *Journalism Quarterly*, pages 415–433, 1953.
- [296] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL <https://doi.org/10.1145/1390156.1390294>.
- [297] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. Xlm-e: Cross-lingual language model pre-training via electra, 2022. URL <https://arxiv.org/abs/2106.16138>.
- [298] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL <https://aclanthology.org/2021.naacl-main.280/>.
- [299] Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages, 2021. URL <https://arxiv.org/abs/2007.15960>.
- [300] Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment objectives for multilingual bidirectional encoders, 2021. URL <https://arxiv.org/abs/2010.07972>.
- [301] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. Improving pretrained cross-lingual language models via self-labeled word alignment. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume*

- 1: *Long Papers*), pages 3418–3430, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.265. URL <https://aclanthology.org/2021.acl-long.265/>.
- [302] Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers, 2022. URL <https://arxiv.org/abs/2205.06266>.
 - [303] Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models, 2024. URL <https://arxiv.org/abs/2401.10440>.
 - [304] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL <https://aclanthology.org/2020.acl-main.421/>.
 - [305] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation, 2013. URL <https://arxiv.org/abs/1309.4168>.
 - [306] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only, 2018. URL <https://arxiv.org/abs/1711.00043>.
 - [307] Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL <https://aclanthology.org/2024.emnlp-main.236/>.
 - [308] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
 - [309] Fei Wang, Kuan-Hao Huang, Kai-Wei Chang, and Muhao Chen. Self-augmentation improves zero-shot cross-lingual transfer. *arXiv preprint arXiv:2309.10891*, 2023.
 - [310] Gyutae Park, Seojin Hwang, Hwanhee Lee, Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao. Low-resource cross-lingual summarization through few-shot learning with large language models. In *Proceedings of LoResMT 2024*, 2024.
 - [311] Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. In *NAACL Long Papers*, 2024.
 - [312] Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, and Rachel Bawden. Cross-lingual strategies for low-resource language modeling: A study on five indic dialects. In *Proceedings of CORIA-TALN 2023*, 2023.
 - [313] Yi-Ting Chiu and Zong-Han Bai. Translation or multilingual retrieval? evaluating cross-lingual search strategies for traditional chinese financial documents. In *FinTech in AI CUP Special Session*, 2025.
 - [314] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. *arXiv preprint arXiv:2101.08370*, 2021.
 - [315] Primer AI. Language agnostic multilingual sentence embedding models for semantic search. 2022.
 - [316] Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. Inducing language-agnostic multilingual representations. *arXiv preprint arXiv:2008.09112*, 2020.
 - [317] Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In Phil Blunsom, Shay Cohen, Paramveer Dhillon, and Percy Liang, editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1521. URL <https://aclanthology.org/W15-1521/>.

- [318] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/gouws15.html>.
- [319] Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data, 2016. URL <https://arxiv.org/abs/1509.07308>.
- [320] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1136. URL <https://aclanthology.org/D16-1136/>.
- [321] Alexis Conneau, Guillaume Lample, Marc’aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.
- [322] Xilun Chen and Claire Cardie. Unsupervised multilingual word embeddings. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1024. URL <https://aclanthology.org/D18-1024/>.
- [323] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL <https://aclanthology.org/P18-1072/>.
- [324] Yuxuan Wang and Lyu Hong. Query encoder distillation via embedding alignment is a strong baseline method to boost dense retriever online efficiency. In Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal, editors, *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustainLP)*, pages 290–298, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.23. URL <https://aclanthology.org/2023.sustainlp-1.23/>.
- [325] Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjanyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. Embeddistill: A geometric knowledge distillation for information retrieval, 2023. URL <https://arxiv.org/abs/2301.12005>.
- [326] Tsegaye Misikir Tashu, Eduard-Raul Kontos, Matthia Sabatelli, and Matias Valdenegro-Toro. Cross-lingual document recommendations with transformer-based representations: Evaluating multilingual models and mapping techniques. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 39–47, Abu Dhabi, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.sumeval-2.4/>.
- [327] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *TACL*, 2019.
- [328] Ruqing Zhao, Jamie Callan, and Luyu Gao. Leveraging sentence-aligned parallel data as supervision signal for neural clir. In *SIGIR*, 2020.
- [329] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *EACL*, 2021.
- [330] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147. URL <https://aclanthology.org/2020.findings-emnlp.147/>.
- [331] Pei-Chi Lo, Yang-Yin Lee, Hsien-Hao Chen, Agus Trisnajaya Kwee, and Ee-Peng Lim. Contrastive learning approach to word-in-context task for low-resource languages. In *Proceedings of the 37th AAAI Workshop on Knowledge Augmented Methods for NLP, AAAI 2023 Workshops*, pages 1–8, Washington, DC, 2023. URL https://ink.library.smu.edu.sg/sis_research/8327.

- [332] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.
- [333] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [334] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.
- [335] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.
- [336] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019.
- [337] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL <https://arxiv.org/abs/2102.05918>.
- [338] Zuohui Fu, Yikun Xian, Shijie Geng, Yingqiang Ge, Yuting Wang, Xin Dong, and Gerard de Melo. Absent: Cross-lingual sentence representation mapping with bidirectional gans. *arXiv preprint arXiv:2001.11121*, 2020.
- [339] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL <https://aclanthology.org/2020.emnlp-main.550/>.
- [340] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=zeFrfgYZ1n>.
- [341] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2021. URL <https://arxiv.org/abs/2104.06967>.
- [342] Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of NAACL-HLT*, 2019.
- [343] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. In *Findings of EMNLP*, pages 1290–1302, 2020.
- [344] Fred Philippy, Siwen Guo, and Shohreh Haddadan. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.323. URL <https://aclanthology.org/2023.acl-long.323/>.
- [345] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [346] Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. Towards multilingual llm evaluation for european languages, 2024. URL <https://arxiv.org/abs/2410.08928>.

- [347] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. Cross-lingual learning-to-rank with shared representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2073. URL <https://aclanthology.org/N18-2073/>.
- [348] Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. Lareqa: Language-agnostic answer retrieval from a multilingual pool, 2020. URL <https://arxiv.org/abs/2004.05484>.
- [349] Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014. URL <https://www.cl.uni-heidelberg.de/~riezler/publications/papers/ACL2014short.pdf>.
- [350] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. Overview of the trec 2023 neuclir track, 2024. URL <https://arxiv.org/abs/2404.08071>.
- [351] Francisco Valentini, Diego Kozłowski, and Vincent Larivière. Clirudit: Cross-lingual information retrieval of scientific documents. *arXiv preprint arXiv:2504.16264*, 2025. URL <https://arxiv.org/abs/2504.16264>. Submitted April 2025.
- [352] Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, Abdul-Hakeem Omotayo, Idris Abdulmumin, Naome A. Etori, Toyib Babatunde Musa, Samuel Fanijo, Oluwabusayo Olufunke Awoyomi, Saheed Abdullahi Salahudeen, Labaran Adamu Mohammed, Daud Olamide Abolade, Falalu Ibrahim Lawan, Maryam Sabo Abubakar, Ruqayya Nasir Iro, Amina Imam Abubakar, Shafie Abdi Mohamed, Hanad Mohamud Mohamed, Tunde Oluwaseyi Ajayi, and Jimmy Lin. Ciral: A test collection for clir evaluations in african languages. *SIGIR '24*, page 293–302, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657884. URL <https://doi.org/10.1145/3626772.3657884>.
- [353] Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. Bordirlines: A dataset for evaluating cross-lingual retrieval augmented generation. *Proceedings of the First Workshop on Advancing NLP for Wikipedia*, Nov 2024.
- [354] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. XQA: A cross-lingual open-domain question answering dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1227. URL <https://aclanthology.org/P19-1227/>.
- [355] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 07 2020. ISSN 2307-387X. doi: 10.1162/tac1_a_00317. URL https://doi.org/10.1162/tac1_a_00317.
- [356] Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering, 2021. URL <https://arxiv.org/abs/2007.15207>.
- [357] Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. Xrag: Cross-lingual retrieval-augmented generation. *arXiv preprint arXiv:2505.10089*, 2025. Published May 2025.
- [358] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL <https://arxiv.org/abs/1611.09268>.
- [359] Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Tammy Li, Runqi Liu, Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness, 2025. URL <https://arxiv.org/abs/2410.01171>.
- [360] Cyril W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield, England, 1962. Also known as the “Cranfield Report”; established the foundation for test-collection based IR evaluation.

- [361] Ellen M. Voorhees. The TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 1999.
- [362] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Draft publicly released, 3rd draft edition, 2025. Draft edition (January 12 2025), available at <https://web.stanford.edu/~jurafsky/slp3/>.
- [363] Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-rewardbench: Evaluating reward models in multilingual settings, 2025. URL <https://arxiv.org/abs/2410.15522>.
- [364] Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Touseh Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. Include: Evaluating multilingual language understanding with regional knowledge, 2024. URL <https://arxiv.org/abs/2411.19799>.
- [365] Shivalika Singh, Angelika Romanou, Clémentine Fourier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2025. URL <https://arxiv.org/abs/2412.03304>.
- [366] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.
- [367] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out*, pages 74–81, 2004.
- [368] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, 2005.
- [369] Maja Popović. chrF: Character n-gram f-score for automatic mt evaluation. In *Proceedings of WMT*, pages 392–395, 2015.
- [370] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, 2006.
- [371] Mikhail Fain, Niall Twomey, and Danushka Bollegala. Backretrieval: An image-pivoted evaluation metric for cross-lingual text representations without parallel corpora. In *Proceedings of SIGIR*, pages 1244–1248, 2021.
- [372] Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. In *Findings of ACL*, pages 123–137, 2024.
- [373] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with bert. In *Proceedings of ICLR*, 2019.
- [374] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation, 2020. URL <https://arxiv.org/abs/2009.09025>.
- [375] Omer Sellam, Noah Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of ACL*, 2020.

- [376] Shuo Sun, Suzanna Sia, and Kevin Duh. Clireval: Evaluating machine translation as a cross-lingual information retrieval task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.
- [377] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 148–151, Los Angeles, June 2010. Association for Computational Linguistics. URL <https://aclanthology.org/W10-0722/>.
- [378] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.16. URL <https://aclanthology.org/2020.eval4nlp-1.16/>.
- [379] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering, 2021. URL <https://arxiv.org/abs/2103.06332>.
- [380] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL <https://arxiv.org/abs/2112.09332>.
- [381] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering, 2023. URL <https://arxiv.org/abs/2305.18201>.
- [382] Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. Modeling exemplification in long-form question answering via retrieval. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.151. URL <https://aclanthology.org/2022.naacl-main.151/>.
- [383] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL <https://aclanthology.org/2022.naacl-main.187/>.
- [384] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL <https://doi.org/10.1145/3209978.3210080>.
- [385] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005. ISSN 0001-4966. doi: 10.1121/1.2016299. URL <https://doi.org/10.1121/1.2016299>.
- [386] Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. Rankgen: Improving text generation with large ranking models, 2022. URL <https://arxiv.org/abs/2205.09726>.
- [387] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS ’21*, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.
- [388] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’98*, page 275–281, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.291008. URL <https://doi.org/10.1145/290941.291008>.

- [389] Statista. Languages most frequently used for web content as of february 2025, by share of websites, 2025. URL <https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/>.
- [390] W3Techs. Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language, 2025. Accessed June 12, 2025.
- [391] Schubert Foo. Retrieval effectiveness of cross language information retrieval search engines. In Chunxiao Xing, Fabio Crestani, and Andreas Rauber, editors, *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*, pages 296–306, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24826-9.
- [392] Nasredine Semmar, Meriama Laib, and Christian Fluhr. A deep linguistic analysis for cross-language information retrieval. In *LREC*, pages 2507–2510, 2006.
- [393] Jiangping Chen and Yu Bao. Cross-language search: The case of google language tools. *First Monday*, 14(3), Feb. 2009. doi: 10.5210/fm.v14i3.2335. URL <https://firstmonday.org/ojs/index.php/fm/article/view/2335>.
- [394] Amin Rahmani and Mohammad Reza Falahati Qadimi Fumani. Adapting google translate for english–persian cross-lingual information retrieval in medical domain. In *International Conference on Information Technology*, 2017.
- [395] Vijay Kumar Sharma, Namita Mittal, and Ankit Vidyarthi. Semantic morphological variant selection and translation disambiguation for cross-lingual information retrieval. *Multimedia Tools and Applications*, 2023.
- [396] Andreea Iana, Goran Glavaš, and Heiko Paulheim. Mind your language: A multilingual dataset for cross-lingual news recommendation. *arXiv preprint arXiv:2403.17876*, 2024.
- [397] Fritz Maurer and Sebastian Voigt. Examining multilingual embedding models cross-lingually through adversarial german–french news search. *arXiv preprint arXiv:2502.08638*, 2025.
- [398] Iknor Singh, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. Breaking language barriers with mmtweets: Advancing cross-lingual debunked narrative retrieval for fact-checking. *arXiv preprint arXiv:2308.05680*, 2023. Introduces a cross-lingual tweet–debunk retrieval dataset.
- [399] Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Maria Bielikova. Multilingual previously fact-checked claim retrieval. In *EMNLP 2023*, pages 16477–16500, 2023. doi: 10.18653/v1/2023.emnlp-main.1027. URL <https://aclanthology.org/2023.emnlp-main.1027>.
- [400] Dorian Quelle, Calvin Yixiang Cheng, Alexandre Bovet, and Scott A. Hale. Lost in translation: Using global fact-checks to measure multilingual misinformation prevalence, spread, and evolution. *EPJ Data Science*, 14(1):22, 2025. doi: 10.1140/epjds/s13688-025-00520-6. Analyzes 264K fact-checks in 95 languages with multilingual embeddings.
- [401] Aryan Singhal, Veronica Shao, Gary Sun, Ryan Ding, Jonathan Lu, and Kevin Zhu. A comparative study of translation bias and accuracy in multilingual large language models for cross-language claim verification. *arXiv preprint arXiv:2410.10303*, 2024. Systematically evaluates translation bias across 15 languages using XFACT.
- [402] Fedor Vitiugin and Carlos Castillo. Cross-lingual query-based summarization of crisis-related social media: An abstractive approach using transformers. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22)*, pages 21–31, 2022. doi: 10.1145/3511095.3531279. Evaluated on five disasters in 10 languages:contentReference[oaicite:1]index=1.
- [403] Cinthia Sánchez, Hernan Sarmiento, Andrés Abeliuk, Jorge Pérez, and Barbara Poblete. Cross-lingual and cross-domain crisis classification for low-resource scenarios. *arXiv preprint arXiv:2209.02139*, 2022. Achieved 80
- [404] Rabindra Lamsal, Maria Rodriguez Read, and Shanika Karunasekera. Semantically enriched cross-lingual sentence embeddings for crisis-related social media texts. *arXiv preprint arXiv:2403.16614*, 2024. Multilingual encoder across 50 languages improving crisis embedding tasks:contentReference[oaicite:3]index=3.
- [405] Fedor Vitiugin and Carlos Castillo. Cross-language classification of crisis-related tweets. *ICWSM Workshop*, 2022. XLM-R fine-tuned English→Arabic crisis tweet classification :contentReference[oaicite:5]index=5.

- [406] Kai Yin, Xiangjue Dong, Chengkai Liu, Lipai Huang, Yiming Xiao, Zhewei Liu, Ali Mostafavi, and James Caverlee. Disastir: A comprehensive information retrieval benchmark for disaster management. *arXiv preprint arXiv:2505.15856*, 2025. Introduced 301 crisis IR tasks across 16 languages :contentReference[oaicite:4]index=4.
- [407] Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Pande. Using deep neural networks to translate multilingual threat intelligence. In *Proceedings of the IEEE Intelligence and Security Informatics Conference (ISI)*, November 2018. Neural pipeline translating Russian-language threat intelligence into English.
- [408] Alex Markov and Mark Last. Identification of terrorist web sites with cross-lingual classification tools. In *Fighting Terror in Cyberspace*, volume 65 of *Studies in Computational Intelligence*, page 117–? Springer, 2005. Early cross-lingual classification for security monitoring.
- [409] Wenya Zhu, Xiaoyu Lv, Baosong Yang, Yinghua Zhang, Yong Xu, Linlong Xu, Yinfu Feng, Haibo Zhang, Qing Da, Anxiang Zeng, and Ronghua Chen. Cross-lingual product retrieval in e-commerce search (clpr-9m). In *Advances in Information Retrieval*, volume 13238 of *Lecture Notes in Computer Science*, pages 441–457. Springer, 2022.
- [410] Xiaoyu Shen, Akari Asai, Bill Byrne, and Adrià de Gispert. xpqa: Cross-lingual product question answering across 12 languages. *arXiv preprint arXiv:2305.09249*, 2023.
- [411] Miriam J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007. doi: 10.1002/asi.20672.
- [412] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018. URL <https://aclanthology.org/P19-1018/>.
- [413] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL <https://aclanthology.org/2021.acl-long.243/>.
- [414] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL <https://aclanthology.org/2023.acl-long.891/>.
- [415] Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing biases and the impact of multilingual training across multiple languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.634. URL <https://aclanthology.org/2023.emnlp-main.634/>.
- [416] Laura Cabello Piqueras and Anders Søgaard. Are pretrained multilingual models equally fair across languages? In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.318/>.
- [417] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models, 2024. URL <https://arxiv.org/abs/2305.14456>.

- [418] Samia Touileb, Lilja Øvrelid, and Erik Velldal. Occupational biases in Norwegian and multilingual language models. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.21. URL <https://aclanthology.org/2022.gebnlp-1.21/>.
- [419] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*, November 2023. ISSN 1396-0466. doi: 10.5210/fm.v28i11.13346. URL <http://dx.doi.org/10.5210/fm.v28i11.13346>.
- [420] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.62. URL <https://aclanthology.org/2022.acl-short.62/>.
- [421] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. BERTScore is unfair: On social bias in language model-based metrics for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.245. URL <https://aclanthology.org/2022.emnlp-main.245/>.
- [422] Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Oluwaseun Eisape, Tanner Spendlove, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. Mmmorrf: Multimodal multilingual modularized reciprocal rank fusion, 2025. URL <https://arxiv.org/abs/2503.20698>.
- [423] Piyush Arora, Dimitar Shterionov, Yasufumi Moriya, Abhishek Kaushik, Daria Dziedzic, and Gareth Jones. An investigative study of multi-modal cross-lingual retrieval. In Kathy McKeown, Douglas W. Oard, Elizabeth, and Richard Schwartz, editors, *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 58–67, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-55-9. URL <https://aclanthology.org/2020.clssts-1.10/>.
- [424] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey, 2024. URL <https://arxiv.org/abs/2408.08921>.
- [425] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. Grag: Graph retrieval-augmented generation, 2024. URL <https://arxiv.org/abs/2405.16506>.
- [426] Tyler Thomas Procko and Omar Ochoa. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169, 2024. doi: 10.1109/AIxSET62544.2024.00030.
- [427] Qiwei Peng, Guimin Hu, Yekun Chai, and Anders Søgaard. Debiasing multilingual llms in cross-lingual latent space, 2025. URL <https://arxiv.org/abs/2508.17948>.
- [428] Diana Abagyan, Alejandro R. Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. One tokenizer to rule them all: Emergent language plasticity via multilingual tokenizers, 2025. URL <https://arxiv.org/abs/2506.10766>.
- [429] François Rémy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*, August 2024. Accepted at COLM 2024.