Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval

Gabriel de Jesus
INESC TEC / University of Porto
Porto, Portugal
gabriel.jesus@inesctec.pt

Siddharth AK Singh
IRLab / University of Amsterdam
Amsterdam, Netherlands
s.a.k.singh@uva.nl

Sérgio Nunes
INESC TEC / University of Porto
Porto, Portugal
sergio.nunes@fe.up.pt

Andrew Yates
HLTCOE / Johns Hopkins University
Maryland, United States
andrew.yates@jhu.edu

Abstract

Dense retrieval models are generally trained using supervised learning approaches for representation learning, which require a labeled dataset (i.e., query-document pairs). However, training such models from scratch is not feasible for most languages, particularly under-resourced ones, due to data scarcity and computational constraints. As an alternative, pretrained dense retrieval models can be fine-tuned for specific downstream tasks or applied directly in zero-shot settings. Given the lack of labeled data for Tetun and the fact that existing dense retrieval models do not currently support the language, this study investigates their application in zero-shot, out-of-distribution scenarios. We adapted these models to Tetun documents, producing zero-shot embeddings, to evaluate their performance across various document representations and retrieval strategies for the ad-hoc text retrieval task. The results show that most pretrained monolingual dense retrieval models outperformed their multilingual counterparts in various configurations. Given the lack of dense retrieval models specialized for Tetun, we combine Hiemstra LM with ColBERTv2 in a hybrid strategy, achieving a relative improvement of +2.01% in P@10, +4.24% in MAP@10, and +2.45% in NDCG@10 over the baseline, based on evaluations using 59 queries and up to 2,000 retrieved documents per query. We propose dual tuning parameters for the score fusion approach commonly used in hybrid retrieval and demonstrate that enriching document titles with summaries generated by a large language model (LLM) from the documents' content significantly enhances the performance of hybrid retrieval strategies in Tetun. To support reproducibility, we publicly release the LLM-generated document summaries and run files.

CCS Concepts

• **Information retrieval** → **Low-resource languages; Ranking**.

Keywords

Tetun, Ad-hoc text retrieval, Zero-shot dense retrieval, Hybrid approaches, Large language models

ICTIR '25, July 18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744593

ACM Reference Format:
Gabriel de Jesus, Siddharth AK Singh, Sérgio Nunes, and Andrew Yates. 2025. Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3731120.3744593

1 Introduction

Traditional sparse retrieval models, such as BM25, which rely on lexical matching, have demonstrated strong performance in many information retrieval (IR) systems [8, 28, 35, 41, 43]. However, because these models retrieve and rank documents based on exact term matching, they are prone to vocabulary mismatch and fail to capture semantic relationships between queries and documents [3, 9, 21, 25, 35].

To bridge this gap, transformer-based dense retrieval models, such as BERT [16] and its variants, have been widely used in text retrieval tasks [17, 25, 30, 31, 37, 48, 51]. These models use dual encoders to map queries and documents independently into low-dimensional vector representations, where semantic relevance is estimated based on vector similarity, typically computed using cosine similarity or dot product functions [17, 21, 31].

Training and fine-tuning dense retrieval models for text retrieval remain computationally expensive and typically require large-scale labeled datasets, which are often difficult to obtain for IR tasks [27]. As a promising alternative, zero-shot text retrieval has gained increasing attention in recent years [5, 27, 39, 43, 52]. More recently, researchers have begun exploring the potential of large language models (LLMs) to support retrieval tasks [24, 55].

MS MARCO [2, 36] is a widely used dataset for training and fine-tuning dense retrieval models and generally uses BERT or mBERT as the backbone models for training. Several dense retrieval models have been trained or fine-tuned on this dataset, including mDPR [52], Contriever and mContriever [23], ColBERT [26, 42], mColBERT [4], ColBERT-X [29], ColBERT-XM [32], ANCE [50], and coCondenser [18, 30].

Studies demonstrate that hybrid retrieval—which combines a traditional sparse retriever with a dense retriever—consistently outperforms a sparse retriever alone, regardless of whether dense retrieval models are fine-tuned [19, 25, 30, 33] or used in zero-shot scenarios, both in- and out-of-distributions [5, 52, 54] and in high- and low-resource languages (LRLs) [52–54]. Two hybrid retrieval approaches have been proposed: (1) score fusion through a linear combination of lexical and semantic matching [19, 25, 30, 33, 52, 53] and (2) reciprocal rank fusion (RRF) [5, 6].

Given the promising results of zero-shot out-of-distribution retrieval in both high- and LRLs [5, 52, 54], this work investigates its feasibility for Tetun. Tetun is a LRL not currently supported by existing pretrained dense retrieval models, which makes the setting especially challenging. The availability of a test collection [13] and baseline for the Tetun ad-hoc text retrieval task [12] provides a valuable foundation for this research.

This study aims to explore the application of a zero-shot out-of-distribution approach to Tetun, investigating the following research questions (RQs):

RQ1 How effective are zero-shot monolingual and multilingual dense retrievers when applied to Tetun in out-of-distribution scenarios compared to traditional sparse retrievers?

RQ2 What is the impact of combining sparse and dense retrievers on the zero-shot retrieval effectiveness for Tetun?

RQ3 How does augmenting document titles with LLM-generated summaries impact zero-shot retrieval effectiveness in Tetun?

To address these RQs, we investigate the performance of pretrained dense retrieval models under zero-shot and hybrid retrieval strategies in out-of-distribution scenarios. Given prior findings that classical sparse retrievers in Tetun perform more effectively when using document titles [12], and that dense retrievers generally benefit from richer contextual signals, we hypothesize that augmenting document titles with content summaries and encoding these enriched representations with a pretrained dense retrieval model may lead to improved retrieval effectiveness for Tetun. We use summaries instead of full documents due to the input length limitations of dense retrievers, which make processing long texts impractical. This approach also allows us to explore how well an LLM can generate content summaries in Tetun.

To test this hypothesis, we initially select pretrained dense retrieval models that have demonstrated strong performance in text retrieval tasks. Subsequently, these models are applied to Tetun queries and document titles (step 1 in Figure 1), and to contextual documents constructed by concatenating document titles with summaries generated by an LLM (step 2 in Figure 1), to create zero-shot embeddings. Next, query and document vectors are computed using cosine similarity to determine their relevance. Finally, the results are linearly combined to produce the final retrieval output.

To evaluate the effectiveness of dense retrieval models for Tetun, we compare zero-shot retrieval performance against the established baseline. Specifically, we assess the performance of dense retrievers, hybrid retrieval using document titles, and hybrid retrieval enhanced with contextual information. The results show that hybrid retrieval performs best when document titles are enriched with summaries generated by an LLM from the documents' content, based on evaluations of up to 2,000 retrieved documents per query. The LLM-generated summaries and run files are publicly available under the Creative Commons Attribution Share-Alike license [15].

2 Background and Related Work

Tetun is one of Timor-Leste's official languages [46], spoken by approximately 79% of the country's 1.18 million population, according to the 2015 census report [11]. The recent release of a Tetun test collection for the ad-hoc text retrieval task [13], along with corresponding baseline results [12], provides a foundation for evaluating the performance of dense retrieval models for Tetun in both zero-shot and hybrid configurations.

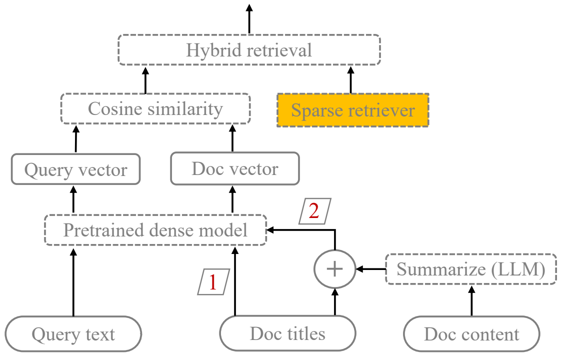

Figure 1: Overview of the Hybrid Retrieval Strategy.

Pretrained dense retrieval models have shown growing potential for improving retrieval effectiveness across languages and retrieval strategies. Early work, such as dense passage retrieval (DPR) [25], demonstrated the effectiveness of dual-encoder architectures in open-domain question answering. This approach was later extended to multilingual settings through mDPR [52], which introduced Mr. TyDI, a multilingual benchmark for monolingual text retrieval with a particular emphasis on LRLs.

Khattab and Zaharia [26] introduced ColBERT, a dense retrieval model that enables fine-grained interaction between queries and documents by comparing token-level representations, allowing for efficient and effective relevance estimation. Its successor, ColBERTv2 [42], improved both accuracy and scalability through reranking and compression mechanisms. The multilingual extensions mColBERT [4], ColBERT-XM [32], and ColBERT-X [29] further adapted the architecture to support retrieval across languages, including low-resource ones. Additionally, models such as ANCE [50], which uses iterative negative sampling, and Contriever and mContriever [23], and coCondenser [18, 30], which are trained without supervision, further advance dense models for text retrieval.

In zero-shot retrieval scenarios, Thakur et al. [43] reported that BM25 maintained strong performance across various corpora, while dense retrievers generally underperformed relative to BM25. Only ColBERT achieved superior results on the BEIR benchmark, outperforming BM25 by an average of +2.5% in NDCG@10 during first-stage retrieval. Similarly, Zhang et al. [52] investigated monolingual ad-hoc text retrieval in non-English languages, focusing on out-of-domain scenarios by training mDPR with mBERT [49] as a replacement for DPR's original encoder. Their findings showed that mDPR did not outperform BM25 when evaluated on the Mr. TyDi benchmark. Building on Mr. TyDi, Zhang et al. [54] introduced MIRACL and reported that dense retrievers, such as mDPR and mContriever, continued to underperform BM25 in zero-shot, out-of-domain settings when evaluated using NDCG@10.

For hybrid retrieval, Karpukhin et al. [25] proposed a score fusion method in which the top 2,000 passages retrieved independently by sparse (BM25) and dense (DPR) retrievers are merged, and their relevance scores are combined using a linear weighted sum. A tuning parameter (λ) was applied to the dense retriever to control its contribution in the fusion process. Experiments on several question-answering (QA) datasets demonstrated that setting $\lambda = 1.1$ on the development set improved end-to-end QA match accuracy. This approach was later adapted by Ma et al. [33] using the same datasets, but the λ tuning parameter was applied to the sparse retriever instead. Its value was optimized via grid search over the range [0, 2] with a step size of 0.05, yielding comparable results.

Gao et al. [19] also experimented with a hybrid retrieval approach for first-stage ad-hoc text retrieval, in which the dense retriever was fine-tuned on the MS MARCO dataset [2]. In this setup, the λ parameter was applied to the sparse retriever during testing, yielding significant improvements over sparse retrieval alone when evaluated on both the MS MARCO and TREC 2019 Deep Learning [7] evaluation query sets. Likewise, Lee et al. [30] evaluated score fusion on MS MARCO passage [36] and TREC Robust04 [47], placing the λ parameter on the dense retriever. The optimal value was selected based on MAP performance on a development set, using a grid search over the range [0, 2] with a step size of 0.1. They found optimal parameter values $\lambda = 1.5$ for MS MARCO and $\lambda = 2.0$ for Robust04. With this configuration, their hybrid approach outperformed existing baselines by combining BM25 with their trained coCondenser model. In a related study, Zhang et al. [54] evaluated hybrid retrieval on the MIRACL datasets, applying α to the sparse retriever and $(1 - \alpha)$ to the dense retriever, with α fixed at 0.5 without tuning. Their results showed that combining mDPR with BM25 outperformed individual retrievers in most languages.

In zero-shot, out-of-domain settings, Zhang et al. [52] evaluated the generalizability and robustness of dense retrievers using a hybrid strategy, where a tuning parameter α was applied to the dense retriever. The value of α was optimized over the range [0, 1] via simple line search on the development set, using MRR@100 as the objective and a step size of 0.01. Combining the top 1,000 results from mDPR and BM25, the hybrid approach significantly outperformed BM25 alone in nine out of eleven languages when evaluated using MRR@100 and Recall@100.

Motivated by these findings, we explore two key directions for extension in Tetun. First, while prior work typically applies a single tuning parameter to the sparse or dense retriever, we assume that assigning distinct parameters to each may more effectively balance their contributions, particularly since sparse and dense models often assign different relevance scores to the same document for a given query. Second, whereas previous studies have primarily focused on passage retrieval and have not incorporated contextual information, we explore the use of enriched document representations. Specifically, we augment document titles with summaries generated by an LLM, adapting pretrained dense retrievers to these contextualized inputs to generate zero-shot embeddings.

3 Collection and Baseline

An overview of the Tetun collection and baseline is provided below.

3.1 Collection Overview

We employed *Labadain-Avaliadór* [13], a Tetun test collection comprising 59 topics, 33,550 documents, and 5,900 query-document relevance judgments (qrels). The documents are sourced from the *Labadain-30k+* dataset [11], with each document containing a title, URL, source, category, publication date, and content. A summary of document titles and content is presented in Table 1.

Table 1: Summary of Document Collection. *Tokens comprise words and numbers.

Description	Total	Min	Max	Avg
#tokens (titles)*	306,840	1	29	9.15
#tokens (content)	11,997,420	2	27,166	357.48

The *Labadain-Avaliadór* [13] is a Tetun test collection with graded relevance levels ranging from 0 (irrelevant) to 3 (highly relevant). It contains between 11 and 99 relevant documents per query, with an average of 36.8. It was developed for the ad-hoc text retrieval task, with query and document relevance evaluated based on topical relevance. Details of *qrels* are presented in Table 2.

Table 2: Summary of Query-Document Relevance Judgments.

Relevance Grade	Total	Proportion
3 - Highly relevant	566	9.59%
2 - Relevant	1,054	17.86%
1 - Marginally relevant	549	9.31%
0 - Irrelevant	3,731	63.24%

3.2 Baseline

We use the baseline reported by de Jesus and Nunes [12], which presents the performance of classical models for ad-hoc text retrieval in Tetun. The results are based on short-text retrieval, in which document titles are indexed after removing hyphens and apostrophes, and model effectiveness is evaluated using the *Labadain-Avaliadór* [13] collection within this setup.

Table 3: Baselines for Ad-hoc Text Retrieval in Tetun.

Model	P@10	MAP@10	NDCG@10
BM25	0.8373	0.2796	0.7347
DFR BM25	**0.8390**	**0.2804**	**0.7356**
Hiemstra LM	0.8305	0.2743	0.7245

The baseline results show that DFR BM25 slightly outperforms both BM25 and the Hiemstra language model (LM) in Precision, MAP, and NDCG at the top ten cutoff, as shown in Table 3. Specifically, DFR BM25 achieves a relative gain of up to 2.22% over Hiemstra LM and a marginal improvement of up to 0.29% over BM25.

4 Dense Retrieval in Tetun

To assess the performance of pretrained dense retrieval models in Tetun, we define four key retrieval settings by leveraging existing pretrained models for text retrieval and adapting them to Tetun documents to produce zero-shot embeddings. We then evaluate how these models perform under different retrieval configurations and examine the impact of contextual augmentation and hybrid retrieval approaches.

The study investigates two key factors: *document representation* and *retrieval strategy*. The former determines whether documents are represented using only titles or are augmented with summaries generated by an LLM to provide additional contextual information. The latter examines whether retrieval relies solely on dense models or integrates dense and sparse models within a hybrid framework. Based on these factors, four retrieval settings are defined:

(1) **Basic Zero-Shot Dense Retrieval:** Pretrained dense retrieval models are applied independently to Tetun queries and document titles to produce zero-shot embeddings, and their retrieval performance is evaluated.

(2) **Contextualized Zero-Shot Dense Retrieval:** Building on (1), document titles are augmented with text summaries generated by an LLM to create zero-shot embeddings. This setting assesses whether enriched contextual information improves zero-shot dense retrieval performance.

(3) **Basic Hybrid Retrieval:** The results from the zero-shot dense retrieval models in (1) are combined with those from sparse retrieval models that use only document titles. This setting examines the effectiveness of hybrid retrieval without additional contextual information.

(4) **Contextualized Hybrid Retrieval:** Building on (3), this strategy combines results from the contextual dense retrieval models in (2) with sparse retrieval outputs to assess the impact of enriched contextual information in a hybrid setup.

We first examine the performance of various pretrained dense retrieval models when adapted to Tetun text documents to produce zero-shot embeddings. We then investigate the impact of incorporating contextual information to enhance these zero-shot embeddings. Next, we evaluate hybrid retrieval techniques by combining sparse and dense retrievers (see Figure 1). Additionally, we compare the performance of monolingual and multilingual variants of the same model. Since the baseline models exhibited similar performance across different cutoff points, all classical models listed in Table 3 were included in the hybrid retrieval experiments.

We focus on pretrained dense retrieval models that have demonstrated strong performance in text retrieval tasks across diverse domains. Selection criteria include the availability of both monolingual and multilingual variants (particularly with support for LRLs), public accessibility, and prior application in hybrid retrieval scenarios. The models selected are DPR [25], mDPR [52], Contriever and mContriever [23], ColBERT [26, 42], and ColBERT-X [29].

5 Dual Parameters for Hybrid Retrieval

In hybrid retrieval, score fusion typically involves combining the scores from both the sparse and dense retrievers through a linear weighted sum, while assigning a tuning parameter to adjust the contribution of either the sparse or dense retriever (referred to as single-parameter tuning). Prior studies have explored various single-parameter tuning strategies, focusing on adjusting either the sparse retriever [19, 33], the dense retriever [25, 30, 52, 53], or both using a single shared parameter [54].

Given that sparse and dense retrievers often assign different relevance scores to the same document for a given query, we propose assigning distinct parameters to each retriever to more effectively balance their contributions. Following established conventions for linear combination [19, 25, 30, 33, 52, 54], we define a dual-parameter tuning approach as follows:

$$\text{Score}(q, d) = \alpha \cdot \text{Score}_{\text{lex}}(q, d) + \beta \cdot \text{Score}_{\text{sem}}(q, d) \tag{1}$$

To identify the optimal combination of α and β, we adopt a strategy inspired by Khramtsova et al. [27] for selecting dense retrievers in zero-shot search scenarios. We perform a grid search over predefined ranges of α and β, evaluating each search step across multiple retrieval strategies and metrics using *qrels* to identify optimal parameter values. At each step of the search, only the parameter values that yield the highest score for each retrieval strategy and metric are recorded for comparison. Given the typically strong performance of sparse baselines, we set α to start from 1.

Prompt 5.1: Details of the LLM Prompt.

User prompt:
Provide a concise contextual description that summarizes the following Tetun document. Do not include any additional text or generate hallucinated content. Do not respond in bullet points, and write exactly one paragraph.

`{document}`

Follow this example:
Input document in Tetun:
`{document_example}`

Expected summary:
`{summary_example}`

System prompt:
You are an expert in Tetun and text understanding. Assume the document is entirely in Tetun and respond using accurate Tetun grammar. Keep your response in Tetun, including correct spelling, punctuation, and other linguistic aspects. Ignore incomplete sentences.

6 Experimental Setup

The experimental configurations include baseline reproduction, document summarization, and retrieval settings, each of which is described in detail below.

6.1 Baseline

To produce baseline runs, we employed the *Labadain-Avaliadór* [13] collection and followed the approach reported by de Jesus and Nunes [12]. PyTerrier [34], a Python API for the Terrier IR platform [38], was employed for indexing, retrieval, and ranking, with the default settings maintained for each model.

During preprocessing, queries and document titles were first lowercased and then tokenized using the Tetun tokenizer [14], followed by the removal of hyphens and apostrophes. The processed documents were then indexed, and retrieval and ranking were performed using BM25 [41], DFR BM25 [1], and Hiemstra LM [40] models to generate run files. These outputs were subsequently evaluated using P@10, MAP@10, and NDCG@10, reproducing the results presented in Table 3.

6.2 Document Summarization and Zero-Shot Embedding Generation

For document summarization, the document content was fed into an LLM to generate a summary. The Haiku variant of the Claude 3 model from Anthropic[1] was used for this task, selected for its cost-effectiveness and its potential to support text summarization in Tetun [10]. The user and system prompts used to instruct the LLM are presented in Prompt 5.1.

These prompts were developed based on preliminary pilot experiments, during which we qualitatively assessed several prompt variants by observing the outputs to identify those that produced coherent and representative summaries of the Tetun documents' content. The LLM was instructed to generate exactly one paragraph of summary for each input document. To guide the model, an example input document in Tetun and its corresponding expected summary were provided in each LLM call. To produce zero-shot embeddings for Tetun, we adapted pretrained dense retrieval models to encode queries and documents. For the documents, we explored two types of representations: titles alone (step 1 in Figure 1) and contextual documents that augment titles with summaries generated from the document content using an LLM (step 2 in Figure 1).

6.3 Zero-Shot and Hybrid Retrieval Strategies

Two strategies were tested and evaluated for zero-shot retrieval. The first strategy involves calculating query-document relevance by computing the cosine similarity between query embeddings and document title embeddings. The second uses contextual embeddings, titles concatenated with content summaries, comparing them to the query embeddings using the same similarity function. The performance of each selected pretrained dense retrieval model, as outlined in Section 4, was evaluated for both strategies.

For zero-shot hybrid retrieval, the document scores produced by each of the zero-shot dense retrievers were linearly combined with the scores from one of the three sparse retrievers (see Subsection 6.1), using the approach proposed in Section 5. To determine the optimal values of α and β, we performed a grid search over $\alpha \in [1, 2]$ and $\beta \in [0, 2]$ with step sizes of 0.01, 0.05, and 0.1, and evaluating performance using P@10, MAP@10, and NDCG@10.

Additionally, the models that outperformed the baseline were further evaluated using the single-parameter linear combination approach and the reciprocal rank fusion (RRF) method to compare their performance. Finally, the performance of monolingual and multilingual dense retrieval models adapted to Tetun was assessed.

7 Results and Analysis

This section presents the experimental results and analysis for six dense retrieval models—DPR, mDPR, Contriever, mContriever, ColBERTv2, and ColBERT-X—evaluated under two configurations: *basic* (using document titles only) and *contextualized* (using document titles enriched with LLM-generated summaries). Performance is evaluated across two retrieval strategies: zero-shot retrieval, in which Tetun documents are adapted to generate zero-shot embeddings, and hybrid retrieval, which combines sparse and dense signals through linear score fusion and reciprocal rank fusion, using three evaluation metrics: P@10, MAP@10, and NDCG@10.

7.1 Zero-Shot Retrieval

Figure 2 presents the relative performance gains of zero-shot dense retrieval models over the baseline sparse retrievers for *basic* and *contextualized* settings. Overall, all dense retrieval models underperform the sparse baseline in zero-shot Tetun text retrieval. Moreover, their performance generally degrades further when contextual information is introduced, as indicated by the orange bars in Figure 2.

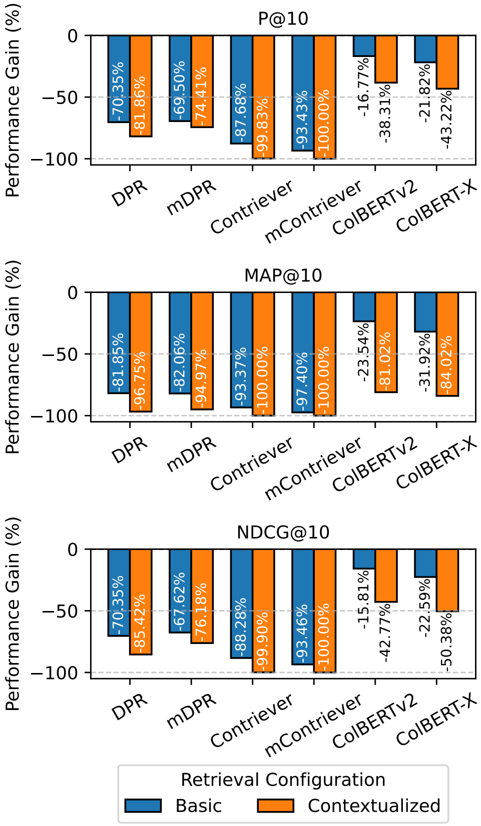

Figure 2: Relative Performance Gains of Zero-Shot Dense Retrieval over the Baseline.

Among the evaluated models, Contriever and mContriever perform the weakest across all metrics, with most of the configurations showing performance losses reaching −100% when contextual information is introduced. DPR and mDPR also underperform consistently relative to the sparse baseline, with further degradation of up to -15% when contextualized inputs are used. While ColBERTv2 and ColBERT-X do not surpass the sparse baseline, they achieve the

[1] https://www.anthropic.com

strongest performance among the dense retrievers under the basic configuration. However, their effectiveness declines considerably with the introduction of enriched document representations, with performance drops of up to -50%.

When comparing each monolingual dense retrieval model to its multilingual counterpart, mixed trends are observed. None of the multilingual models were trained on Tetun. For the DPR family, the multilingual variant exhibits a smaller performance drop than its monolingual counterpart. In contrast, for both Contriever and ColBERT variants, the monolingual models demonstrate lower performance loss compared to their multilingual counterparts.

7.2 Hybrid Retrieval

In hybrid retrieval, we conducted a grid search over $\alpha \in [1, 2]$ and $\beta \in [0, 2]$ and the optimal parameters identified were $\alpha = 1.0$ and $\beta = 0.3$ for the *basic* configuration, and $\alpha = 1.1$ and $\beta = 0.8$ for the *contextualized* configuration, with a step size of 0.1 used in both cases. Results for each configuration are presented below.

7.2.1 Basic Hybrid Retrieval. The relative performance gains over the DFR BM25 baseline for the *basic* hybrid retrieval strategy are presented in Figure 3. Overall, most combinations of dense retrievers with Hiemstra LM yield the smallest performance losses relative to the baseline. Notably, the ColBERT-X + Hiemstra LM combination slightly outperforms the baseline, achieving a +0.25% relative performance gain in MAP@10 (middle bar graph in Figure 3). In contrast, the Contriever family continues to exhibit the largest performance degradation, with losses reaching up to -60%. Among the evaluated models, DPR, ColBERTv2, and ColBERT-X demonstrate minimal performance loss when combined with sparse retrievers, each under 5% compared to the baseline. In contrast, mDPR shows moderate degradation, with losses ranging from -13% to -20%. Contriever continues to perform significantly worse, with drops between -40% and -52%, while mContriever records the highest performance degradation, ranging from -47% to -60%.

When comparing *basic* hybrid configuration results to their zero-shot retrieval counterparts shown in Figure 2, significant performance improvements are observed. For instance, at MAP@10, and when combined with Hiemstra LM, DPR improves from a −81.85% loss in the zero-shot setting to just −0.64%; Contriever improves from −93.37% to −52.00%; and most notably, ColBERT-X shifts from a −84.02% performance drop in zero-shot to a +0.25% gain.

Regarding the comparison between monolingual and multilingual dense retrieval models, nearly all monolingual models outperform their multilingual counterparts when combined with sparse retrievers and are generally close to the baseline performance. An exception is ColBERT-X combined with Hiemstra LM at MAP@10, where it slightly exceeds the baseline with a relative performance gain of +0.25%. ColBERTv2, meanwhile, exhibits consistent performance across all combinations and evaluation metrics.

7.2.2 Contextualized Hybrid Retrieval. Based on the optimal parameter values identified through grid search in the *basic* hybrid retrieval setup, where $\alpha = 1.0$ indicated that no adjustment to the sparse retriever and optimal performance was achieved using a single tuning parameter $\beta = 0.3$ for the dense retriever, we extended the evaluation of single-parameter tuning to the *contextualized* configuration. In this regard, we assess both single-parameter and dual-parameter fusion strategies: single-parameter tuning applies either α to the sparse retriever or β to the dense retriever, while the dual-parameter setup assigns independent weights to both. Moreover, we evaluate RRF as an alternative fusion method. The results for contextualized hybrid retrieval are presented in Table 4. Given our objective to demonstrate that balancing signal fusion from both dense and sparse retrievers, particularly when incorporating contextual information, yields better performance than one-sided tuning, we begin by presenting the results from the dual-parameter configuration, followed by the single-parameter settings.

Dual Parameters. The results of dual-parameter tuning, with optimal parameter values of $\alpha = 1.1$ and $\beta = 0.8$, are presented in Table 4. These results show that combining ColBERTv2 with each of the sparse retrievers consistently outperforms the baseline across all evaluation metrics. The best performance is achieved by the ColBERTv2 + Hiemstra LM combination, yielding a relative improvement of +2.01% in P@10, +4.24% in MAP@10, and +4.24% in NDCG@10 over the baseline, based on evaluations of up to 2,000 retrieved documents per query. These improvements are statistically significant, as confirmed by a paired t-test ($p < 0.05$). Scores exceeding the baseline are shown in bold, with the best results highlighted in green. Conversely, other dense retrievers exhibit less consistent behavior when introducing contextual information. For instance, when combined with Hiemstra LM in MAP@10, ColBERT-X incurs only a marginal additional loss of -0.14 percentage points relative to its *basic* hybrid configuration. In contrast, DPR, mDPR, Contriever, and mContriever experience substantially greater degradations, with additional losses ranging from -4.67 to -25.57 percentage points.

Single Parameter. To evaluate the effectiveness of single-parameter tuning, we selected the models that outperformed the baseline in the dual-parameter setting while preserving the optimal parameter values identified for the sparse and dense retrievers. Thus, ColBERTv2 and ColBERT-X were chosen for this analysis, with the corresponding results reported in Table 4. Overall, tuning a single parameter for the ColBERTv2-based model, either in the sparse or dense retriever, continues to yield strong performance relative to the baseline, except for the ColBERTv2 + DFR BM25 combination when the parameter is applied to the sparse retriever. In contrast, ColBERT–X–based combinations show less consistent behavior. For example, the previously observed MAP@10 improvement when combined with Hiemstra LM drops to -0.50% in the single-parameter setting. In general, applying the tuning parameter to the dense retriever proves more effective than tuning the sparse retriever, except in the case of ColBERTv2 combined with BM25. Notably, tuning the sparse retriever in the ColBERTv2 + BM25 combination outperforms its dual-parameter counterpart across all metrics, achieving relative gains of up to +0.93 percentage points.

Reciprocal Rank Fusion (RRF). To further assess the effectiveness of the hybrid retrieval strategy for Tetun ad-hoc text retrieval, we evaluated the performance of ColBERTv2 and ColBERT-X using the RRF method, as shown in Table 4. While the combination of ColBERTv2 + Hiemstra LM shows a slight advantage, RRF does not outperform the baseline and offers limited benefit for Tetun.

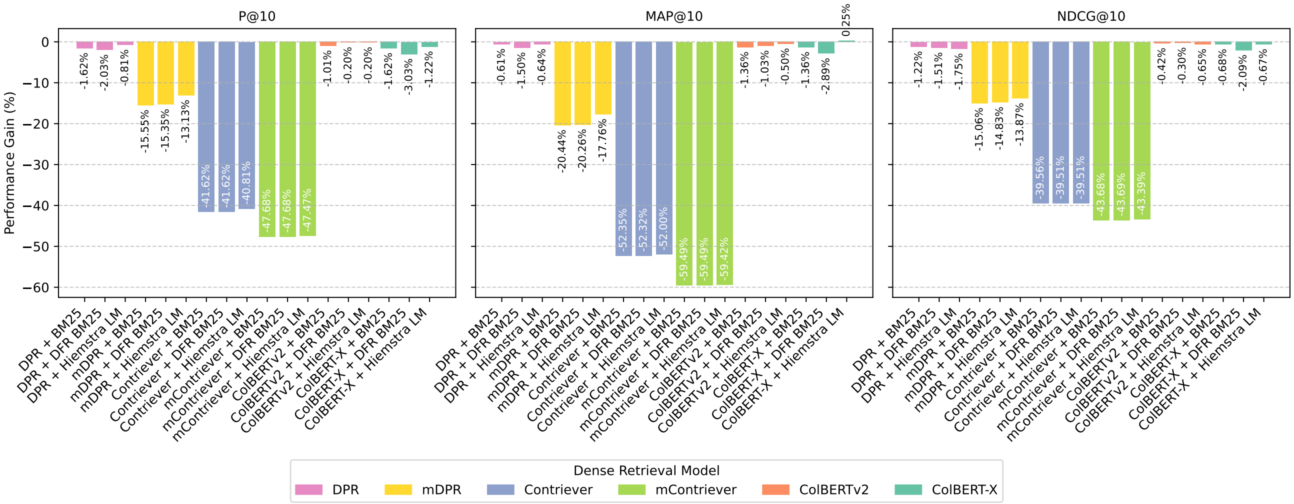

Figure 3: Relative Performance Gains of Hybrid Retrieval over the Baseline under the *Basic* Configuration ($\alpha = 1.0$, $\beta = 0.3$).

In the *contextualized* hybrid configuration, nearly all monolingual models consistently outperform their multilingual counterparts. The only exception is the mDPR model, which achieves better results across all evaluation metrics in the dual-parameter setting.

8 Discussion

The results of this study are consistent with prior studies [43, 52, 54], showing that all evaluated dense retrieval models underperform relative to the sparse baseline (DFR BM25) in zero-shot scenarios, regardless of whether the models are monolingual or multilingual, with monolingual models exhibiting a slight advantage over their multilingual counterparts. These outcomes reaffirm the limitations of dense retrieval models in generalizing across domains and languages outside their training distribution, particularly in LRLs, and highlight the robustness of traditional sparse models.

A contextualization strategy introduced in this study—enriching document representations by concatenating document titles with LLM-generated content summaries—does not improve performance in zero-shot, out-of-distribution scenarios. On the contrary, it leads to further performance degradation, as illustrated in Figure 2 (orange bars). These findings suggest that current dense retrieval models are limited in their ability to leverage additional contextual signals in unfamiliar domains and may be particularly sensitive to noisy, redundant, or domain-mismatched inputs.

Hybrid retrieval, in contrast, when dense retrievers were combined with sparse models under the *basic* hybrid configuration (as shown in Figure 3), all evaluated combinations showed notable gains relative to their zero-shot counterparts in Figure 2. The hybrid configuration combining ColBERT-X with Hiemstra LM was particularly promising, achieving near-baseline performance and, in one case, slightly surpassing it by +0.25% in MAP@10.

While these findings are broadly consistent with prior work, they differ in input data: previous studies predominantly employ passages for training or fine-tuning, whereas this study uses document titles, which are significantly shorter and often less informative. This distinction likely contributes to the lower standalone effectiveness of dense retrievers in our setting. Nevertheless, the results underscore the effectiveness of combining lexical and semantic signals, even when the dense component performs poorly in isolation.

Further improvements were observed in the *contextualized* hybrid setting using dual-parameter tuning, as presented in Table 4. ColBERTv2, when combined with Hiemstra LM and tuned with $\alpha = 1.1$ and $\beta = 0.8$, yielded statistically significant gains across all evaluation metrics (P@10, MAP@10, and NDCG@10), with relative improvements of up to +4.24%. This suggests that proper calibration of the balance between sparse and dense signals is essential to maximizing hybrid retrieval effectiveness. Interestingly, these improvements were not uniformly observed across all models, reinforcing that architectural differences among dense retrievers impact their ability to leverage contextual and fused signals.

Single-parameter tuning experiments further reveal that strong performance can still be achieved when tuning only one component. For ColBERTv2, sparse-side tuning in combination with BM25 not only performed competitively but also outperformed the model's dual-parameter configuration by up to +0.93 percentage points. This finding suggests that in practical scenarios, particularly those that rely on BM25 as the sparse retriever, a simpler fusion strategy focused on the sparse side may offer an effective and efficient alternative to more complex dual-parameter approaches.

The performance of reciprocal rank fusion (RRF), a rank-based hybridization technique, was less consistent. Although some benefit was observed for ColBERTv2 combined with Hiemstra LM, RRF generally did not outperform score fusion methods and often underperformed relative to the baseline. This outcome indicates that, in the context of Tetun ad-hoc text retrieval, score-based hybridization provides finer control over the influence of each retriever and may be more suitable than rank-only fusion strategies.

The comparison between monolingual and multilingual dense retrievers revealed a notable performance gap favoring monolingual models, particularly in hybrid configurations. This suggests that

Table 4: Performance of Contextualized Hybrid Retrieval in Tetun. Scores in bold indicate improvement over the baseline. Values highlighted with a *green* background represent the best performance compared to the baseline (paired t-test, $p < 0.05$). The symbol † indicates that the model outperformed its counterpart (monolingual vs. multilingual). Rows highlighted with a *grey* background correspond to multilingual models.

	P@10	MAP@10	NDCG@10
Baseline			
DFR BM25	0.8390	0.2804	0.7356
Contextualized hybrid retrieval ($\alpha = 1.1$ and $\beta = 0.8$, top 2k results)			
DPR + BM25	0.6864 (-18.19%)	0.2079 (-25.86%)	0.6161 (-16.25%)
DPR + DFR BM25	0.6864 (-18.19%)	0.2083 (-25.71%)	0.6169 (-16.14%)
DPR + Hiemstra LM	0.6847 (-18.39%)	0.2061 (-26.50%)	0.6102 (-17.05%)
mDPR + BM25	0.7068 (-15.76%)†	0.2226 (-20.61%)†	0.6236 (-15.23%)†
mDPR + DFR BM25	0.7102 (-15.35%)†	0.2234 (-20.33%)†	0.6262 (-14.87%)†
mDPR + Hiemstra LM	0.7288 (-13.13%)†	0.2305 (-17.80%)†	0.6334 (-13.89%)†
Contriever + BM25	0.3169 (-62.23%)†	0.0791 (-71.79%)†	0.2938 (-60.06%)†
Contriever + DFR BM25	0.3169 (-62.23%)†	0.0791 (-71.79%)†	0.2947 (-59.94%)†
Contriever + Hiemstra LM	0.3220 (-61.62%)†	0.0796 (-71.61%)†	0.2985 (-59.42%)†
mContriever + BM25	0.1814 (-78.38%)	0.0408 (-85.45%)	0.1936 (-73.68%)
mContriever + DFR BM25	0.1814 (-78.38%)	0.0409 (-85.41%)	0.1938 (-73.65%)
mContriever + Hiemstra LM	0.1847 (-77.99%)	0.0421 (-84.99%)	0.1967 (-73.26%)
ColBERTv2 + BM25	**0.8441 (+0.61%)**†	**0.2862 (+2.07%)**†	**0.7453 (+1.32%)**†
ColBERTv2 + DFR BM25	**0.8458 (+0.81%)**†	**0.2858 (+1.93%)**†	**0.7480 (+1.69%)**†
ColBERTv2 + Hiemstra LM	**0.8559 (+2.01%)** †	**0.2923 (+4.24%)** †	**0.7536 (+2.45)** †
ColBERT-X + BM25	0.8305 (-1.01%)	0.2765 (-1.39%)	0.7335 (-0.29%)
ColBERT-X + DFR BM25	0.8322 (-0.81%)	0.2779 (-0.89%)	0.7308 (-0.65%)
ColBERT-X + Hiemstra LM	0.8356 (-0.41%)	**0.2807 (+0.11%)**	0.7303 (-0.72%)
Contextualized hybrid retrieval (single parameter, $\alpha = 1.1$, top 2k results)			
ColBERTv2 + BM25	**0.8508 (+1.41%)**†	**0.2888 (+3.00%)**†	**0.7493 (+1.86%)**†
ColBERTv2 + DFR BM25	0.8373 (-0.20%)†	**0.2831 (+0.96%)**†	**0.7438 (+1.11%)**†
ColBERTv2 + Hiemstra LM	**0.8475 (+1.01%)**†	**0.2891 (+3.10%)**†	**0.7504 (+2.01%)**†
ColBERT-X + BM25	0.8322 (-0.81%)	0.2771 (-1.18%)	0.7334 (-0.30%)
ColBERT-X + DFR BM25	0.8339 (-0.61%)	0.2771 (-1.18%)	0.7325 (-0.42%)
ColBERT-X + Hiemstra LM	0.8356 (-0.41%)	0.2790 (-0.50%)	0.7304 (-0.71%)
Contextualized hybrid retrieval (single parameter, $\beta = 0.8$, top 2k results)			
ColBERTv2 + BM25	**0.8458 (+0.81%)**†	**0.2866 (+2.21%)**†	**0.7467 (+1.51%)**†
ColBERTv2 + DFR BM25	**0.8441 (+0.61%)**†	**0.2856 (+1.85%)**†	**0.7476 (+1.63%)**†
ColBERTv2 + Hiemstra LM	**0.8542 (+1.81%)**†	**0.2917 (+4.03%)**†	**0.7530 (+2.37%)**†
ColBERT-X + BM25	0.8322 (-0.81%)	0.2775 (-1.03%)	0.7344 (-0.16%)
ColBERT-X + DFR BM25	0.8339 (-0.61%)	0.2774 (-1.07%)	0.7327 (-0.39%)
ColBERT-X + Hiemstra LM	0.8373 (-0.20%)	0.2804 (+0.00%)	0.7305 (-0.69%)
Reciprocal Rank Fusion			
ColBERTv2 + BM25	0.7949 (-5.26%)†	0.2641 (-5.81%)†	0.7164 (-2.61%)†
ColBERTv2 + DFR BM25	0.7966 (-5.05%)†	0.2642 (-5.78%)†	0.7166 (-2.58%)†
ColBERTv2 + Hiemstra LM	0.8220 (-2.03%)†	0.2738 (-2.35%)†	0.7316 (-0.54%)†
ColBERT-X + BM25	0.7898 (-5.86%)	0.2575 (-8.17%)	0.6965 (-5.32%)
ColBERT-X + DFR BM25	0.7898 (-5.86%)	0.2577 (-8.10%)	0.6969 (-5.26%)
ColBERT-X + Hiemstra LM	0.8034 (-4.24%)	0.2604 (-7.13%)	0.6985 (-5.04%)

multilingual models may not generalize Tetun effectively without further adaptation. An exception was ColBERT-X, which, despite its multilingual nature, demonstrated strong stability and, in certain configurations, nearly matched or exceeded the performance of monolingual ColBERTv2. This suggests that multilingual dense retrievers can be competitive when paired with late interaction mechanisms and carefully tuned hybrid strategies.

The relationship between language-specific features, pretraining data, and model architecture is a key factor influencing retrieval performance in Tetun. Tetun's typological status—as a language with extensive lexical borrowing from Portuguese [11, 20, 22, 44, 45]—is reflected in the query and document content, with an average of approximately 47% of tokens in queries and 30% in contextual documents being Portuguese loanwords, as shown in Table 5. Examples of such loanwords in Tetun include *"ekonomia"* (from *"economia"*, meaning "economy") and *"progresu"* (from *"progresso"*, meaning "progress"). This linguistic overlap may help explain the robust performance of ColBERT-X, although it does not fully account for the consistently higher performance of monolingual ColBERTv2.

Table 5: Statistics for Queries and Documents Containing Portuguese Loanwords. Loanwords were identified by matching each preprocessed word against the dictionary published by Greksáková [20]. *Tokens include words and numbers.

Description	Min	Max	Avg
Total tokens (queries)*	3	5	3.46
Total loanwords (queries)	0	3	1.61
Total tokens (contextual documents)	4	335	122.87
Total loanwords (contextual documents)	0	99	36.57

One plausible explanation—aligned with the broader discussion on the balance between sparse and dense signals—is that both the queries and the documents favor lexical matching. From this perspective, the competitive performance of the ColBERT architecture may stem from its late-interaction mechanism, which preserves token-level signals. Likewise, the superior performance of the monolingual ColBERT variant might be attributed to the high proportion of Portuguese (Latinate) loanwords in Tetun, which aligns well with ColBERTv2's pretraining on English, a language that also contains substantial Latinate influence in its lexicon and morphology.

Overall, these findings emphasize the nuanced interplay between retrieval architecture, fusion strategy, and contextual input. While dense retrievers face substantial challenges in zero-shot Tetun ad-hoc text retrieval, their performance can be substantially enhanced through hybrid approaches, architectural choices, and tuning parameter balancing. The results reinforce attention to hybrid retrieval design and suggest contextual adaptation in LRL scenarios.

9 Conclusions and Future Work

This study investigates the retrieval effectiveness of adapting pretrained dense retrievers to Tetun, a low-resource language (LRL), under zero-shot and hybrid retrieval scenarios in out-of-distribution settings. In the zero-shot setting, in response to RQ1, our results confirm that dense retrievers, both monolingual and multilingual, consistently underperform relative to the sparse retriever baseline. This performance gap highlights the limitations of current dense retrieval models in generalizing to LRLs not seen during training. While monolingual dense models slightly outperform their multilingual counterparts, none surpass the baseline, reaffirming the robustness of traditional sparse retrieval models in low-resource, out-of-domain contexts and underscoring the challenge of adapting dense models for such scenarios.

In hybrid retrieval strategies, addressing RQ2 and RQ3, our findings show that hybrid approaches are generally less effective when using short text inputs, such as document titles. However, performance improves substantially when these titles are enriched with contextual information—augmenting titles with summaries generated by a large language model (LLM) before encoding.

The best performance is achieved with the ColBERTv2 + Hiemstra LM combination. A relative improvement in MAP@10 of +3.10% is observed when assigning $\alpha = 1.1$ to the sparse retriever, and +4.03% when assigning $\beta = 0.8$ to the dense retriever. The highest improvement of +4.24% in MAP@10 is achieved when applying both parameters ($\alpha = 1.1$, $\beta = 0.8$), indicating the effectiveness of the dual-tuning approach in hybrid retrieval. These improvements highlight the effectiveness of both single and dual-parameter strategies when contextual information is incorporated. This supports a contextualization approach that is particularly beneficial in scenarios where passage-level datasets are unavailable, a common challenge in under-resourced language settings. The enhanced retrieval effectiveness observed through contextual enrichment confirms our hypothesis that augmenting short-text representations with LLM-generated summaries can significantly improve performance in zero-shot retrieval.

This work contributes two key methodological innovations. First, we propose a dual-parameter tuning strategy, grounded in the intuition that sparse and dense retrievers often assign different relevance scores to the same document. By weighting each retriever independently, this approach enables more effective balancing of their complementary signals, improving performance. Second, we introduce a contextualized retrieval strategy for zero-shot scenarios, where LLM-generated summaries are concatenated with titles to enrich representations before generating dense embeddings.

To support reproducibility and foster future research, we publicly release two resources: contextual document summaries generated by an LLM and all associated retrieval run files [15]. These resources provide a foundation for advancing research in Tetun information retrieval and more broadly across other multilingual LRL contexts.

As part of future work, we aim to evaluate the generalizability of the proposed methods across a broader range of multilingual scenarios, with a particular focus on monolingual ad-hoc text retrieval tasks, and the relationship between linguistic characteristics of Tetun and model performance. Moreover, we plan to investigate the lightweight representation learning techniques and the integration of advanced reranking methods, including the use of LLMs as post-retrieval rerankers, to assess their performance in Tetun.

10 Acknowledgments

This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia I.P., under the support UID/50014/2023 (DOI: 10.54499/UID/50014/2023). Gabriel de Jesus is supported by the Ph.D. studentship grant SFRH/BD/151437/2021, also from FCT (DOI: 10.54499/SFRH/BD/151437/2021). This work was partly supported by project VI.Vidi.223.166 of the NWO Talent Programme which is (partly) financed by the Dutch Research Council NWO).

References

[1] Gianni Amati and C. J. van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (2002), 357–389. https://doi.org/10.1145/582415.582416

[2] 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. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL] https://arxiv.org/abs/1611.09268

[3] Adam L. Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu O. Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, Emmanuel J. Yannakoudakis, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong (Eds.). ACM, 192–199. https://doi.org/10.1145/345508.345576

[4] Luiz Henrique Bonifacio, Israel Campiotti, Roberto A. Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. *CoRR* abs/2108.13897 (2021). arXiv:2108.13897 https://arxiv.org/abs/2108.13897

[5] Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-Domain Semantics to the Rescue! Zero-Shot Hybrid Retrieval Models. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 95–110. https://doi.org/10.1007/978-3-030-99736-6_7

[6] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel (Eds.). ACM, 758–759. https://doi.org/10.1145/1571941.1572114

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2019. Overview of the TREC 2019 Deep Learning Track. In *Proceedings of The Twenty-Eighth Text REtrieval Conference (TREC 2019)*. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.DL.pdf

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the TREC 2023 Deep Learning Track. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023 (NIST Special Publication, Vol. 500-xxx)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/Overview_deep.pdf

[9] W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines - Information Retrieval in Practice.* Pearson Education. http://www.search-engines-book.com/

[10] Gabriel de Jesus and Sérgio Nunes. 2024. Exploring Large Language Models for Relevance Judgments in Tetun. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024), co-located with the 10th International Conference on Online Publishing (SIGIR 2024)*, C. Siro, M. Aliannejadi, H.A. Rahmani, N. Craswell, C.L.A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz (Eds.), Vol. 3752. Washington D.C., USA, 19–30. https://ceur-ws.org/Vol-3752/

[11] Gabriel de Jesus and Sérgio Nunes. 2024. Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, Maite Melero, Sakriani Sakti, and Claudia Soria (Eds.). ELRA and ICCL, Torino, Italia, 177–188. https://aclanthology.org/2024.sigul-1.22

[12] Gabriel de Jesus and Sérgio Nunes. 2025. Establishing a Foundation for Tetun Text Ad-Hoc Retrieval: Stemming, Indexing, Retrieval, and Ranking. arXiv:2412.11758 [cs.IR] https://arxiv.org/abs/2412.11758

[13] Gabriel de Jesus and Sérgio Nunes. 2025. Labadain-Avaliadór: A Test Collection for Tetun Ad-hoc Text Retrieval Task [Dataset]. https://doi.org/10.25747/2k6s-e518

[14] Gabriel de Jesus and Sérgio Sobral Nunes. 2024. Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a Tetun Text Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, 4368–4380. https://aclanthology.org/2024.lrec-main.390

[15] Gabriel de Jesus, Siddharth AK Singh, Sérgio Nunes, and Andrew Yates. 2025. Labadain-ZSRunS: A Sparse and Zero-Shot Dense Retrieval Runs with LLM-Generated Summaries for Tetun Ad-Hoc Text Retrieval [Data set]. INESC TEC. https://doi.org/10.25747/rfzx-m945

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[17] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, and Jiafeng Guo. 2022. Pre-training Methods in Information Retrieval. *Found. Trends Inf. Retr.* 16, 3 (2022), 178–317. https://doi.org/10.1561/1500000100

[18] Luyu Gao and Jamie Callan. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2843–2853. https://doi.org/10.18653/V1/2022.ACL-LONG.203

[19] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement Lexical Retrieval Model with Semantic Residual Embeddings. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12656)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 146–160. https://doi.org/10.1007/978-3-030-72113-8_10

[20] Zuzana Greksáková. 2018. *Tetun in Timor-Leste: The role of language contact in its development.* Ph. D. Dissertation. Universidade de Coimbra, Portugal. http://hdl.handle.net/10316/80665

[21] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Trans. Inf. Syst.* 40, 4 (2022), 66:1–66:42. https://doi.org/10.1145/3486250

[22] John Hajek and Catharina Williams van Klinken. 2019. Language Contact and Gender in Tetun Dili: What Happens When Austronesian Meets Romance? *Oceanic Linguistics* 58 (06 2019), 59–91. https://doi.org/10.1353/ol.2019.0003

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

[24] Nour Jedidi, Yung-Sung Chuang, Leslie Shing, and James R. Glass. 2024. Zero-Shot Dense Retrieval with Embeddings from Relevance Feedback. *CoRR* abs/2410.21242 (2024). https://doi.org/10.48550/ARXIV.2410.21242 arXiv:2410.21242

[25] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550

[26] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[27] Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Xi Wang, and Guido Zuccon. 2023. Selecting which Dense Retriever to use for Zero-Shot Search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Beijing, China) *(SIGIR-AP '23)*. Association for Computing Machinery, New York, NY, USA, 223–233. https://doi.org/10.1145/3624918.3625330

[28] Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach. *CoRR* abs/2010.01195 (2020). arXiv:2010.01195 https://arxiv.org/abs/2010.01195

[29] Dawn J. Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2023. Neural Approaches to Multilingual Information Retrieval. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 521–536. https://doi.org/10.1007/978-3-031-28244-7_33

[30] Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi, and Sunghyun Park. 2023. On Complementarity Objectives for Hybrid Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 13357–13368. https://doi.org/10.18653/V1/2023.ACL-LONG.746

[31] Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond.* Morgan & Claypool Publishers. https://doi.org/10.2200/S01123ED1V01Y202108HLT053

[32] Antoine Louis, Vageesh Kumar Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2025. ColBERT-XM: A Modular Multi-Vector Representation Model for Zero-Shot Multilingual Information Retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, 4370–4383. https://aclanthology.org/2025.coling-main.295/

[33] Xueguang Ma, Kai Sun, Ronak Pradeep, Minghan Li, and Jimmy Lin. 2022. Another Look at DPR: Reproduction of Training and Replication of Retrieval. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 613–626. https://doi.org/10.1007/978-3-030-99736-6_41

[34] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich (Eds.). ACM, 161–168. https://doi.org/10.1145/3409256.3409829

[35] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 1 (2018), 1–126. https://doi.org/10.1561/1500000061

[36] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[37] Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 708–718. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.63

[38] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier Information Retrieval Platform. In *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings (Lecture Notes in Computer Science, Vol. 3408)*, David E. Losada and Juan M. Fernández-Luna (Eds.). Springer, 517–519. https://doi.org/10.1007/978-3-540-31865-1_37

[39] Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. A Thorough Examination on Zero-shot Dense Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 15783–15796. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.1057

[40] Stephen Robertson and Djoerd Hiemstra. 2001. Language models and probability of relevance. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, J. Callan, B. Croft, and J. Lafferty (Eds.). Carnegie Mellon University, United States, 21–25.

[41] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[42] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 3715–3734. https://doi.org/10.18653/V1/2022.NAACL-MAIN.272

[43] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html

[44] Catharina Williams van Klinken and John Hajek. 2018. Language contact and functional expansion in Tetun Dili: The evolution of a new press register. *Multilingua* 37 (2018), 613–647. https://doi.org/10.1515/multi-2017-0109

[45] Catharina Williams van Klinken, John Hajek, and Rachel Nordlingerh. 2002. *Tetun Dili: a grammar of an East Timorese language.* Pacific Linguistics, Canberra, Australia. https://doi.org/10.15144/PL-528

[46] Pedro Carlos Bacelar de Vasconcelos, Andreia Sofia Pinto Oliveira, Ricardo Sousa da Cunha, Andreia Rute da Silva Baptista, Alexandre Corte-Real de Araújo, Benedita McCrorie Graça Moura, Bernardo Almeida, Cláudio Ximenes, Fernando Conde Monteiro, Henrique Curado, et al. 2011. Constituição Anotada da República Democrática de Timor-Leste. http://hdl.handle.net/10400.22/4008

[47] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf

[48] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md. Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Comput. Surv.* 56, 7 (2024), 185:1–185:33. https://doi.org/10.1145/3648471

[49] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 38–45. https://doi.org/10.18653/V1/2020.EMNLP-DEMOS.6

[50] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[51] Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple Applications of BERT for Ad Hoc Document Retrieval. *CoRR* abs/1903.10972 (2019). arXiv:1903.10972 http://arxiv.org/abs/1903.10972

[52] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multilingual Benchmark for Dense Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 127–137. https://doi.org/10.18653/v1/2021.mrl-1.12

[53] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2024. Toward Best Practices for Training Multilingual Dense Retrieval Models. *ACM Trans. Inf. Syst.* 42, 2 (2024), 39:1–39:33. https://doi.org/10.1145/3613447

[54] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Trans. Assoc. Comput. Linguistics* 11 (2023), 1114–1131. https://doi.org/10.1162/TACL_A_00595

[55] Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2024. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 4375–4391. https://aclanthology.org/2024.emnlp-main.250