A Zero-Shot Monolingual Dual Stage Information Retrieval System for Spanish Biomedical Systematic Literature Reviews

Regina Ofori-Boateng[1], **Magaly Aceves-Martins**[2], **Nirmalie Wirantuga**[1], and **Carlos Francisco Moreno-García**[1]

[1]School of Computing, Robert Gordon University, {r.ofori-boateng, n.wiratunga, c.moreno-garcia}@rgu.ac.uk
[2]The Rowett Institute, University of Aberdeen, {magaly.aceves@abdn.ac.uk}

Abstract

Systematic Reviews (SRs) are foundational in healthcare for synthesising evidence to inform clinical practices. Traditionally skewed towards English-language databases, SRs often exclude significant research in other languages, leading to potential biases. This study addresses this gap by focusing on Spanish, a language notably underrepresented in SRs. We present a foundational zero-shot dual information retrieval (IR) baseline system, integrating traditional retrieval methods with pre-trained language models and cross-attention re-rankers for enhanced accuracy in Spanish biomedical literature retrieval. Utilising the LILACS database, known for its comprehensive coverage of Latin American and Caribbean biomedical literature, we evaluate the approach with three real-life case studies in Spanish SRs. The findings demonstrate the system's efficacy and underscore the importance of query formulation. This study contributes to the field of IR by promoting language inclusivity and supports the development of more comprehensive and globally representative healthcare guidelines.

1 Introduction

Systematic reviews (SRs) are crucial in healthcare as they represent the gold standard for synthesising evidence to inform clinical guidelines, policies and practices (O'Mara-Eves et al., 2015). Their comprehensive nature ensures that all relevant studies on a particular topic are evaluated, providing a holistic understanding of evidence (Shojania et al., 2010). However, existing studies have shown that most SRs are skewed towards English databases, excluding databases in Languages other than English (LoE), even when the SRs are geographically specific (Walpole, 2019; Neimann Rasmussen and Montgomery, 2018). Yet, the results of these studies inform guidelines and practices internationally. Additionally, it has been proven that LoE papers offer insights, context, or results not available in English-language databases (Machado, 2016). These insights alter the direction of association or the effect of meta-analyses (Walpole, 2019). For instance, a reported SR exploring risk factors for youth violence found that if LoE databases were not incorporated, 15-30% of the included studies would have been omitted (Shenderovich et al., 2016).

Thus, to ensure that SRs are comprehensive, which is one of its hallmarks, all evidence must be considered regardless of language or geographical origin. Yet, one major problem reported as to why these LoE databases are not included in SR is the navigation and search skills required (Walpole, 2019). These databases often differ from standard English ones, requiring researchers to possess specialised search skills (Flemming et al., 2018). Thus, valuable research will not be noticed without proper Information Retrieval (IR) techniques, compromising the comprehensiveness of clinical decision-making. In this study, we aim to use a *Spanish SR database* as a case study. It must be noted that Spanish (the fourth most spoken language in the world) is one of the most neglected yet relevant languages used for SRs (Walpole, 2019; Aceves-Martins et al., 2022). We present a state-of-the-art (SOTA) baseline to serve as a foundation for future research. As such, the present work introduces a foundational benchmark for retrieving Spanish biomedical literature and provides insights into, 1) querying the database with search queries obtained from the research question/title as done in English SRs database search[1] (Jin et al., 2019), and 2) using synonyms of the outcome of the SR as the search strings. For instance, in one of the SRs to be used as a case study on mental health **Salud mental** (Aceves-Martins et al., 2022), we explore the effect of using queries from the research question itself (e.g. *¿Cuál es la asociación entre la obesidad o el sobrepeso Y los problemas de salud*

[1]https://training.cochrane.org/handbook/current

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3725–3736

mental entre los niños Y adolescentes mexicanos?) or the use of the synonyms outcome of the SR study at hand (*depresion Y ninos Y Mexico*) provided by human experts. The proposed method consists of a zero-shot dual information retrieval system: a hybrid first-stage retrieval that combines encodings from term-based and a Pre-trained Language Model (PLM), and a re-ranker model to further refine the outputs from the first stage. The model is trained with Spanish abstracts retrieved from the LILACS database[2] and evaluated on real-life SR case studies. We opted to use LILACS because it stands as the premier and most extensive index for biomedical scientific and technical literature in Latin America and the Caribbean.

Existing studies have shown the potential of a two-stage system in IR (retrieval and re-ranking) (Nogueira and Cho, 2020; Ma et al., 2020a; Glass et al., 2022). Yet, these methods are mainly skewed towards English biomedical literature and databases that require labelled or synthetic training data. The main contributions of this paper are 1) compared to existing IR dual-stage retrieval methods skewed towards English biomedical literature, we present a zero-shot dual hybrid first-stage retrieval for Spanish SR case studies to serve as a benchmark for future study, 2) we generate and train the dual IR system with the original Spanish title as question/query to form the question-abstract pair rather than synthetically generating these questions, considering the particularities of how literature is written in the Spanish-speaking world and our available computational resources, and 3) we experiment the model with four SOTA Spanish PLM and report and discuss the result as well provide statistical insights into query formation.

2 Related Studies

Numerous methods have been proposed for IR tasks, including first-stage retrieval techniques, re-rankers, and hybrid approaches. To begin, the first-stage retrieval has primarily relied on classic term-based probabilistic models like Best Match (BM25) (Robertson and Zaragoza, 2009), known for their high efficiency and effectiveness, even with very large document collections (Chakraborty et al., 2023). With the advancement in neural networks, dense retrieval methods have been proposed to encode documents and queries calculating relevance through a similarity function (Palangi et al., 2016; Cohen et al., 2018; Reimers and Gurevych, 2019; Karpukhin et al., 2020). Recent research in text re-ranking has increasingly focused on transformer-based models (Vaswani et al., 2017). (Nogueira et al., 2019) and (Gao et al., 2021) proposed using BERT-based cross-attention method to capture interactions between queries and passages. Other studies have proposed encoder-decoder PLM, such as T5, for text binary ranking purposes (Pradeep et al., 2021; Zhuang et al., 2023; Raffel et al., 2020). Also, (Muennighoff, 2022) introduced SGPT for ranking documents. Furthermore, (Sachan et al., 2022) proposed a zero-shot UPR PLM to re-rank passages directly.

Recently, studies have focused on hybrid models for IR. Authors in (Neji et al., 2021; Kuzi et al., 2020; Ma et al., 2020a) proposed the use of a combination of BM25 and BERT (Devlin et al., 2019) PLM for first-stage retrieval. (Nogueira and Cho, 2020) presented the use of a sparse traditional BM25 model as the retriever and monoBERT and duoBERT as the re-ranker model. (Ma et al., 2020b) proposed using a hybrid first retrieval with BERT and BM25, a cross-attention model trained with PubMed abstract on the BioASQ[3] test dataset. Similarly, (Lu et al., 2022b) also proposed the use of a zero-shot first-stage model for the BioASQ challenge with a dual BERT encoder and dual T5-re-ranker model. Finally, (Glass et al., 2022) presented Re2G, a method that combines BM25, GPT-3 and T5 in a dual-stage information system. Though these methods have shown promising results, a major setback is all the studies focus on English datasets and some require synthetically generated (document-query pair) for training data which can be computationally costly and affect the authenticity of the data. As such, inspired by the work done by (Lu et al., 2022b) this study seeks to expand this research frontier by investigating a dual IR retrieval system specifically tailored for Spanish biomedical literature in a zero-shot learning environment, addressing the critical.

3 Methodology

3.1 Question-Abstract Generation for training

To help us obtain the queries for training the model without manually labelling or synthetically generating like (Ma et al., 2021) on their PubMed

[2]https://lilacs.bvsalud.org/es/

[3]http://bioasq.org/

Table 1: Statistics of the Spanish literature data collected from LILACS

Statistics	2009-2012	2013-2015	2016-2019	2020-2023
Total Obtained	47388	35939	50426	40663
Avg abstract length	180.49	187.61	193.15	180.56
Avg title length	25.304	26.46	27.13	25.31

abstract, we followed the same method in creating the PubMedQA (Jin et al., 2019) benchmark dataset. This helped us enhance the authenticity of the content and reflect real-world information. The PubMedQA dataset is an English biomedical Q&A of abstracts obtained from the PubMed database, where the questions are derived from the titles of PubMed articles. Additionally, this approach was used to reduce computation resources. Thus, we converted the titles of biomedical abstracts indexed and collected from the LILACS database to questions/queries. We collected Spanish abstracts from 2009 to 2023. Table 2 summarises the statistics of the unlabelled dataset for training the zero-shot hybrid first-stage retrieval in the benchmark dual information retrieval system.

To ensure the high quality of our training data, we developed an in-house cleaning pipeline. Upon collecting the Spanish abstracts, we first fixed **Mojibake**[4], an encoding issue by recording misinterpreted characters back to their original byte form and then decoding those bytes, alongside the use of regular expressions. In addition, duplicates and null abstracts were removed. The abstracts and questions were tokenised to pass the dataset into the term-based model, and stopwords were removed using the Spacy Spanish[5] packages.

Figure 1 and Algorithm 1 explain the methodology used. The model encompasses a dual-stage system: an initial retrieval system and a re-ranking system. In this work, the initial (or first-stage) retrieval system is a hybrid system which combines a strong, traditional sparse IR baseline (BM25) for keyword matching and a dual SOTA-dense PLM (four Spanish-based PLM described in Section 3.2) for contextual/semantic understanding. For the second/re-ranker stage, we use a cross-attention of the four individual PLMs. The re-ranker undergoes training with the subset of candidates that have been initially retrieved in the first stage. The following subsection describes these stages in detail. Thus, given the set of abstract A and a query q, the dual systems comprises a retrieval function $r : \{(q, a_j) \mid a_j \in A\} \rightarrow \mathbb{R}$, where $r(q, a_j)$ assigns a relevance score to each abstract a_j with respect to the query q. The re-ranker function is defined as $f : \{(q, a'_k) \mid a'_k \in A'\} \rightarrow \mathbb{R}$, where $A' \subset A$ is the subset of documents deemed potentially relevant by the retriever, $f(q, a'_k)$ and reorders the documents in A' based on their calculated relevance to the query q.

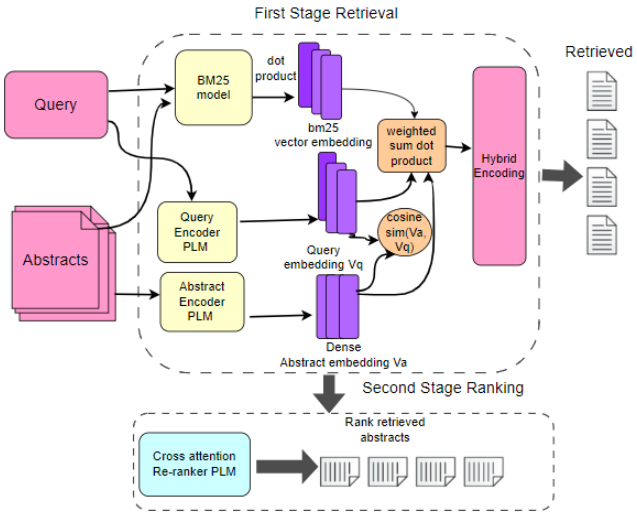

Figure 1: Overview of the dual-stage IR system

3.2 Pre-trained Language models

To build the dual-stage IR system, we explored four PLMs that support Spanish literature: multilingual BERT (mBERT) (Devlin et al., 2019), XLM-R Galén (Cross-lingual Language Model - RoBERTa), a continual pre-training version of the XLM-R model (Conneau et al., 2020) trained on Galén Oncology corpus (Lopez-Garcia et al., 2021), Spanish-BERT (BETO) (Cañete et al., 2023), and bsc-bio-ehr-es, a PLM trained on Electronic Health Records (EHR) in Spanish (Carrino et al., 2022). For simplicity, we will refer to the bsc-bio-ehr-es PLM as Bsc-EHR. The choice of these models was guided by their proven efficacy in recent comparative studies for biomedical tasks in Spanish

[4]https://en.wikipedia.org/wiki/Mojibake
[5]https://spacy.io/models/es

(Aracena and Dunstan, 2023), with each model bringing a distinct linguistic and contextual understanding critical for our research. mBERT is trained on texts from 104 different languages incorporating a diverse, multilingual WordPiece vocabulary comprising around 110,000 subwords and possesses about 177M trainable parameters. XLM-R Galén is a continuous pre-training of the XLM-R on medical datasets. XLM-R, which was pre-trained on a common crawl corpus encompassing 100 languages. It utilises an extensive multilingual SentencePiece vocabulary, which includes approximately 250,000 subwords. BETO is a model pre-trained solely based on Spanish texts and encompasses approximately 110 million trainable parameters. Lastly, the Bsc-EHR model was trained on a different biomedical corpus comprising a total of 514,000 documents and 95 million tokens with a Longformer (Lopez-Garcia et al., 2021) backbone.

3.3 Stage 1: Initial Retrieval

3.3.1 Traditional sparse (BM25) model

The BM25 is a keyword algorithm used widely in IR systems that ranks a set of documents based on the query terms appearing in each document. It is based on the probabilistic retrieval framework, which aims to estimate the probability that a given document is relevant to the user's query using lexical matching. Using inverted indexes for inference, the BM25 uses lexical sparsity per document to optimise retrieval speed and memory usage. The core of the BM25 algorithm is its scoring function, which is used to estimate the relevance of a given document to a query. To implement the BM25 model, we pass the preprocessed queries and abstracts described in Section 3.1. Similar to these studies, (Lu et al., 2022a; Ma et al., 2020a), both A and d are represented in a vector space. Each q, is encoded into a vector, denoted as $\vec{q}_{\text{BM25}}$, in a high-dimensional binary vector where the number of dimensions is equal to the total number of unique words in the entire abstract represented by $|A|$. In this vector, each dimension corresponds to a word from $|A|$ and is set to 1 if the word is present in the query else 0. In a manner akin to q, each abstract is also transformed into a sparse vector, $\vec{a}_{\text{BM25}}$. Given a term a_i in the abstract A, the value of each vector component is given by Equation 3.3.1.

$$\vec{a_i}_{\text{BM25}} = \frac{\text{IDF}(a_i)\text{cnt}(a_i, A)(k+1)}{\text{cnt}(a_i, A) + k(1 - b + b\frac{m}{\text{avgdl}})} \quad (1)$$

where a_i refers to an individual word in A, $\text{cnt}(a_i, A)$ is the frequency of word a_i in the A, k and b are hyper-parameters to influence of term frequency and abstract length provided in Section 4, $\text{IDF}(a_i)$ is a score reflecting the importance of word a_i across all documents (inverse document frequency), m is the number of words in A, and avgdl is the average length of the abstract. We represent the BM25 model in a vector-based representation, though not typical of standard BM25 implementation, to make it compatible with the output PLM to combine the lexical strength from the BM25 and semantic understanding from the PLM. During inference, the BM25 score is computed for each A in the corpus with respect to the query using the formula above. The documents are then ranked based on these scores, and the top-scoring documents are returned as the most relevant to the query.

3.3.2 Dense Retrieval Model (Dual Encoder)

Compared to the BM25 model, the dual encoder PLM captures the semantic meaning between the query and abstracts. The primary goal of training the dual encoder is to be able to generate high-quality embedding for question-abstract pair (Dong et al., 2022). Thus, these embeddings capture the semantic meaning of the texts. The PLM dual encoder is trained to maximise the similarity between the related abstract-question pairs and minimise it for incorrect pairs achieved using a contrastive loss (Chopra et al., 2005). The model computes relevance scores based on the similarity between these vectors by encoding queries and abstracts into a dense vector space. This similarity score indicates how relevant a document is to a given query. The following Section describes the training of the dual encoder in detail.

3.3.3 Training the Dual Encoder

To train the dual encoder to capture semantic relations, discern between semantically related and unrelated pairs, and learn discriminative features, we generated negative samples (abstract-question pairs) using in-batch negatives from the training data over traditional negative sampling methods due to efficiency and the diversity of negatives it will provide within each batch and enhance generalisability (Xiong et al., 2020). Thus, instead of explicitly sampling separate negative examples for a given query in the batch, all documents paired

Algorithm 1 Overview of the dual-stage IR system

Require: Biomedical Query q, Abstract Set A
Ensure: Ranked list of documents

Step 1: BM25 Model
Preprocess and tokenize q.
Divide A into batches and preprocess and tokenize each batch.
Encode q into a binary vector $\vec{q}_{\text{BM25}}$, with each dimension representing a unique word in A. Set dimensions to 1 for words present in q, else 0.
for each batch of abstracts in A **do**
    Calculate BM25 scores for all abstracts in the batch relative to q in Equation 3.3.1.
    Save the BM25 scores
end for
Step 2: Dual Encoder Model (mBERT, BETO, XLM-R Galén and bsc-bio-ehr-es PMLs
Initialise the PLM model for query and document encoding
Generate negative samples using in-batch negatives
Encode tokenised input abstract a and query q into dense vectors $v_a = E_A(a)$ and $v_q = E_Q(q)$
for each document d in D **do**
    Encode Q and d into dense vectors using BERT
    Compute cosine similarity score between encoded Q and d
end for
Step 3: Hybrid Model
Set interpolation parameter λ
for each document d in D **do**
    Compute hybrid score as $\lambda \cdot \langle q_{\text{bm25}}, d_{\text{bm25}} \rangle + \langle q_{\text{de}}, d_{\text{de}} \rangle$
end for
Sort documents in D based on hybrid scores
Step 4: Cross-Attention Reranker
Select top N documents based on hybrid scores
for each selected document d **do**
    Combine Q and d as input to a PLM-based cross-attention model
    Rerank d based on the output of the cross-attention model
end for
return Reranked list of documents

with other queries in the same batch are treated as negatives for that query. This was to enable the dual encoder to map abstracts and questions in the vector space and understand which mappings are meaningful (positive) and which are not (negative) and for training efficiency. The dual encoder model comprises two encoding components, both instantiated from the individual PLM in Section 3.2. We use a dual encoder model with a siamese architecture, using two identical encoders (sharing weights and structure) to encode both the query and the abstract. Each encoder transforms the input text data into dense vector representations in a shared embedding space. Thus, mathematically, given the tokenised input abstract a and a query q, their respective vector representations are $v_a = E_A(a)$ and $v_q = E_Q(q)$, where E_A and E_Q represent the encoding functions for abstracts and questions. The goal of the dual encoder is to learn embedding such that the distance (or dissimilarity) between embedding of similar pairs is minimised, while that for dissimilar pairs is maximised. As such, we used the cosine contrastive loss L to include a margin that defines how far apart the dissimilar pairs should be. Thus, the loss function that governed the training of the dual encoder given a batch of N pairs of abstracts and questions, with each pair labelled as positive ($y = 1$) or negative ($y = -1$) is:

$$L = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} 1 - \cos(v_{a_i}, v_{q_i}), & \text{if } y_i = 1 \\ \max(0, \cos(v_{a_i}, v_{q_i}) - m), & \text{if } y_i = -1 \end{cases} \tag{2}$$

where $\cos(\cdot)$ denotes the cosine similarity function, and v_{a_i}, v_{q_i} are the vector representations of the i-th abstract and question, respectively and m is the hyperparameter margin.

3.3.4 Hybrid First-Stage: Concatenation of Sparse and PLM Encoding

To leverage the strengths of both lexical term-based features from BM25 and semantic similarities from the dense retrieval model, we concatenate the vector scores from the sparse, BM25 and dense encodings (de). This process results in a hybrid vector for each query and document, encapsulating both the term-based features from BM25 and semantic features from the neural model. Following (Ma et al., 2021), we introduce an interpolation hyperparameter (λ), set to 0.5. This hyperparameter balances the influence of the BM25 and neural components in the hybrid model, allowing for an adjustable emphasis between lexical and semantic matching. The

Table 2: Description of SR evaluation dataset

Dataset	Research Question (Spanish)	Papers retrieved (LILACS)	Total Papers	Total Included	Included papers (LILACS)
Oral Health (SB)	¿Existe asociación entre obesidad o sobrepeso y mala salud bucal en niños y adolescentes mexicanos?	73	9828	18	3
Mental Health (SM)	¿Cuál es la asociación entre la obesidad o el sobrepeso y los problemas de salud mental entre los niños y adolescentes mexicanos?	35	1074	16	6
Obesity Prevention (PO)	¿Cuál es la efectividad de las intervenciones de prevención de la obesidad entre los niños mexicanos?	20	9828	29	2

similarity score in the hybrid model is thus calculated as a weighted sum of the dot product of BM25 and dual encoder similarities. Mathematically, the hybrid similarity score is expressed as:

$$\text{sim}(q_{\text{hyb}}, a_{\text{hyb}}) = \lambda\langle q_{\text{bm25}}, a_{\text{bm25}}\rangle + \langle q_{\text{de}}, a_{\text{de}}\rangle \quad (3)$$

where $\langle\cdot,\cdot\rangle$ indicates the dot product. Here, q_{hyb} and a_{hyb} are the hybrid encodings of the query and document, respectively; q_{bm25} and a_{bm25} are their BM25 encodings; and q_{de} and a_{de} are their dual encoder embeddings. The interpolation hyperparameter λ provides flexibility in adjusting the relative contribution of lexical and semantic information in the hybrid model.

3.4 Stage 2: Cross-attention Re-ranker

To enhance the precision of the hybrid first-stage retrieval system, we implement a re-ranker model using a cross-attention (Zhang et al., 2022). This was to focus on the nuanced alignment between the content of a document and the user's query intent, which is pivotal for assessing the relevance of a document to a specific query (Gao et al., 2021). We utilised four PLMs to construct our cross-attention model. The input format for the re-ranker is structured as follows: [CLS] question [SEP] abstract [SEP]. In processing these inputs, the PLMs enable each word (or token) to "attend" to every other word in the sentence. Such an arrangement allows for a deep contextual understanding, ensuring each word is interpreted with the entire input sequence. This process leads to the generation of enriched and contextually aware representations. The re-ranker is trained using the listwise method similar to that described in (Lu et al., 2022b). The objective is to enable the model to proficiently rank relevant documents higher than non-relevant ones for a given query. We compile a list comprising one relevant (positive) example and M non-relevant (negative) examples for each question. In our training, we utilise a list size of 50, which includes one positive and 49 negative examples with a binary relevance (relevant or not relevant). We explored both a ranking-based loss, a pairwise hinge and a binary cross-entropy loss, yet we selected the latter as it performed better.

4 Experimental Setup

4.1 Evaluation Dataset

To evaluate the performance of the investigated model, we used three biomedical human annotated Spanish SR case studies: **SB: Oral health and obesity in Mexico (salud bucal y obesidad en México)**(Aceves-Martins et al., 2022), **SM: Mental health and obesity in Mexico (salud mental y obesidad en México)**(Godina-Flores et al., 2022), and **PO: Obesity prevention in Mexican children (prevención de obesidad en niños mexicanos**(Aceves-Martins et al., 2021). We selected this domain especially in children because it has been reported as a complex issue and a severe health problem (Mercado-Mercado, 2023). This evaluation dataset is part of an ongoing Mexican project, Childhood Obesity in Mexico (COMO)[6]. The three SRs were queried from the LILACS data source, plus others such as EMBASE and Medline. These were manually annotated by the expert systematic reviewers of the COMO project. The human annotators are experienced systematic reviewers within the health domain and have experience reviewing in the Cochrane Collaboration[7]. Although, we do acknowledge the existence of MESINESP2[8], a Spanish medical semantic indexing dataset which focuses on assigning relevant medical concepts to medical texts/abstracts to facilitate IR, this work does not focus on semantic medical indexing. Instead, this research focuses on IR for Spanish SRs. Thus, to the best of our knowledge, we did not find any publicly available SR evaluation biomedical IR dataset curated and annotated from Spanish databases. As such, we make our evaluation dataset available together with the detailed descriptions of the potential records

[6]https://www.comoprojectmx.com/
[7]https://www.cochrane.org/
[8]https://temu.bsc.es/mesinesp2/

that were found within each of the three SRs such as their DOIs, titles, descriptors, among others[9]. Beyond presenting a benchmark for future studies, we seek to investigate the impact of using queries directly from the research question/title/objective as done in English biomedical studies (Jin et al., 2019), compared to using the study's outcome provided by human experts. For instance, in *SB*, the research question used in the study was *Is there an association between obesity/overweight and poor oral health in Mexican children and adolescents? (¿Existe asociación entre obesidad o sobrepeso y mala salud bucal en niños y adolescentes mexicanos?)*, but the refined version provided by the SR researchers contained more specific terms of the outcome of the study, such as "cavities AND children AND Mexico" ("caries Y niños Y México"). As such, we obtained 8 queries for both the SB and SM datasets and 15 queries for the PO dataset. We provide a summary of the three SR datasets in Table 2.

4.2 Nearest Neighbour Search

During the evaluation, we explored the concept of nearest neighbour search, a method for identifying the closest or most similar data points in a dataset (abstract) to a specific query point. We chose this method over the approximate nearest neighbours after careful experimentation and consideration of the trade-offs between precision/recall and efficiency, along with the advantages and disadvantages of each approach. Consequently, we opted for the nearest neighbour search to evaluate the BM25 algorithm, since it is represented in high-dimensional (sparse) vectors, the dual encoder and the hybrid retrieval model on the SR test set.

4.3 Training and Hyper-parameter setup

All code was written in Python. The Pytorch framework[10] and Hugging Face hubs were used for loading and training the PLMs[11]. Unlike traditional machine learning models, BM25 does not require a learning phase but fine-tuning of its hyper-parameters to enhance retrieval accuracy. To find the optimal hyper-parameters, we do not perform the traditional train/test split of the Spanish (abstract-question pairs). Instead, we find the optimal hyper-parameters through grid search from ranges of k = [**1.5**, 1.6, 1.7, 1.9, 2.0] and b = [0.5, 0.65, 0.75, **0.85**, 0.95] by using different subsets of data based on their varying lengths and topics. To find the optimal hyper-parameters for the dual encoder and cross-attention models, we further split the training set into a validation set (80% train, 20% validation) and find the best values through a series of experiments optimised using AdamW. For the dual encoder, the optimal values are: Batch size: $[8, 16, \mathbf{32}, 64, 256]$, warm up step: $[0, \mathbf{500}, 1000]$, weight decay: $[0, 0.1, \mathbf{0.01}, 0.001]$, learning rate: $[\mathbf{1e^{-4}}, 3e^{-5}, 1e^{-5}]$, and epochs: [3, 5, **10**, 20]. We implemented early stopping as a regularisation to prevent overfitting. Furthermore, we used the same values and method for training the cross-attention re-ranker PLMs. In this case, the optimal values are: Batch size: $[8, \mathbf{16}, 32, 64, 256]$ and weight decay: $[0, \mathbf{0.1}, 0.01, 0.001]$. The experiments were conducted on Google Colab, A100 GPU. Using the best hyper-parameters, we varied three random seeds (10, 42, 50) and averaged the results of the evaluation dataset across three iterations.

4.4 Evaluation Metrics

To evaluate the model, we use standard metrics for IR systems based on the TREC standard evaluation metrics, that is the Mean Average Precision (MAP) and the Normalised Discounted Cumulative Gain (nDCG) for the queries of each SR case study. MAP focuses on measuring the ability of the model to retrieve relevant abstracts, while nDCG evaluates both the relevance of the documents retrieved and their ranking order in the search results, thus a measure of ranking quality.

4.5 Baselines

To evaluate the foundational dual-stage IR model to be built on, we evaluate it against the individual components of the dual-stage model: Strong **BM25** baseline, the individual four dense PLMs, **mBERT, BETO, Bsc-EHR and XLM-R Galén** and the **hybrid combinations** (BM25+mBERT), (BM25+BETO), (BM25+Bsc-EHR) and (BM25+XLM-R Galén) of the PLMs against the dual-stage model.

5 Results and Discussion

Table 3 shows the results obtained by using queries from the original title/research question of the study. A clear observation to be made from the table is that the hybrid models (combination of BM25 and PLMs) generally improved the performance over the single models, indicating the effectiveness

[9] https://github.com/reginaofori/Zero_Shot_IR_Spanish
[10] https://pytorch.org/
[11] https://huggingface.co/

Table 3: Summary of results from querying using research title/question of SR evaluation study. ↑ represents the increment of the best results compared to the strongest baseline (either PLM/BM25) within each category approximated to 3 d.p. The **bold** values represent the highest values in each category/PLM type and **bold** values also denote the overall best results within each dataset category.

Querying from the research title	**SB**		**SM**		**PO**	
Models	**MAP**	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**
BM25	0.0033	0.0000	0.0000	0.1072	0.0690	0.0984
m-BERT Dual Encoder	0.0000	0.0000	0.0000	0.1061	0.1303↑ 0.028	0.0904
m-BERT Hybrid	0.0137	0.0000	0.0000	0.1186	0.0804	0.1072 ↑ 0.007
m-BERT Hybrid Re-Ranker	0.0200↑ 0.006	0.0000	0.0000	0.1240 ↑ 0.005	0.1020	0.1004
XLM-R Galén Dual Encoder	0.0150	0.0000	0.0040	0.0848	0.1119	0.1071
XLM-R Galén Hybrid	0.0386	0.0198	**0.0264**↑ 0.016	0.1370	0.1121	0.1150
XLM-R Galén Hybrid Re-ranker	**0.0803**↑ 0.042	**0.0251** ↑ 0.005	0.0104	**0.1596** ↑ 0.023	**0.1223** ↑ 0.010	**0.1172**↑ 0.002
Bsc-EHR Dual Encoder	0.0046	0.0000	0.0182 ↑ 0.017	0.1095	0.0629	0.0793
Bsc-EHR Hybrid	0.0060	0.0108	0.0014	0.0758	0.1029	0.0800
Bsc-EHR Hybrid Re-Ranker	0.0629↑ 0.057	0.0182↑ 0.007	0.0014	0.1206↑ 0.011	0.1103 ↑ 0.007	0.0914 ↑ 0.011
BETO Dual Encoder	0.0140	0.0188	0.0051	0.0838	0.0709	0.0994
BETO Hybrid	0.0376	0.0000	0.0094	0.1176	0.0710	0.1161 ↑ 0.002
BETO Hybrid Re-ranker	0.0720 ↑ 0.034	0.0193↑ 0.001	0.0120↑ 0.003	0.1286 ↑ 0.011	0.1151 ↑ 0.044	0.1140

Table 4: Summary of results with expert modification- using refined terms of the study outcome. ↑ represents the increment of the best results compared to the strongest baseline within each category approximated to 3 d.p. The **bold** values represent the highest values in each category/PLM type and **bold** values also denote the overall best results within each dataset category

Synonyms of outcome of the SR study	**SB**		**SM**		**PO**	
Models	**MAP**	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**
BM25	0.1033	0.0567	0.0713	0.2046	0.1720	0.2014
m-BERT Dual Encoder	0.0694	0.0791	0.0603	0.2091	0.1633	0.2402
m-BERT Hybrid	0.1167	0.0722	0.0939	0.2166	0.1834	0.2610 ↑ 0.021
m-BERT Hybrid Re-ranker	0.1233 ↑ 0.007	0.0830 ↑ 0.004	0.0983 ↑ 0.004	0.2309 ↑ 0.014	0.2180 ↑ 0.035	0.2402
XLM-R Galén Dual Encoder	0.1260	0.0708	0.0651	0.1358	0.1629	0.1414
XLM-R Galén Hybrid	0.1296	0.0424	0.0871 ↑ 0.022	0.1696	0.1230	0.1421
XLM-R Galén Hybrid Re-ranker	0.1329 ↑ 0.003	0.0782 ↑ 0.007	0.0614	0.2206 ↑ 0.051	0.1703	0.1514 ↑ 0.009
Bsc-EHR Dual Encoder	0.1167	0.0944	0.1103	0.2216	0.1750	0.2022
Bsc-EHR Hybrid	0.1181	0.1229	**0.1215** ↑ 0.003	0.1879	0.2150	**0.2224**
Bsc-EHR Hybrid Re-ranker	**0.1750** ↑ 0.057	**0.1303** ↑ 0.007	0.1135	**0.2257** ↑ 0.052	**0.2727** ↑ 0.058	**0.2224** ↑ 0.021
BETO Dual Encoder	0.1450 ↑ 0.015	0.0708	0.0614	0.1358	0.1510	0.1514
BETO Galén Hybrid	0.1296	0.0424	0.0782 ↑ 0.016	0.1696	0.1630	0.1703
BETO Galén Hybrid Re-ranker	0.1229	0.0782 ↑ 0.007	0.0614	0.2206 ↑ 0.051	0.1703	0.1514

of combining term-based and semantic approaches. Also, the addition of a re-ranker stage (dual stage) generally improved the results compared to simply using the hybrid models, suggesting the effectiveness of this additional refinement step in the IR process. In summary, the XLM-R Galén hybrid re-ranker model showed promising results across all three SR studies, with the highest scores in both MAP and nDCG metrics. Particularly, it achieved notable results for the SO and PO studies.

Moving on to Table 4, where we used the refined terms of outcomes in queries provided by experts, the retrieval effectiveness is enhanced across all models compared to querying from the research title. For instance, in the previous Table 3, where models such as BM25 had very poor performance, we now see a notable improvement. Another interesting observation is that in Table 3 XLM-R Galén hybrid re-ranker gives the best results, and in the improved one, the Bsc-EHR Hybrid Re-ranker performed better in the various case studies, in particular in SM and PO. This indicates its effectiveness in retrieving and re-ranking potential biomedical abstracts. Nonetheless, there is also a consistent improvement in hybrid models: Across all SR studies, the results obtained for the hybrid models and re-ranker imply their effectiveness in combining term-based retrieval with semantic understanding from PLMs and ranking relevant abstracts. To further provide insights into the comparison between Table 3 and Table 4, Table 5, presents the results of paired t-tests comparing the original terms with refined outcome terms for MAP and nDCG. For datasets that did not meet the assumptions of the paired t-test, as determined by the Shapiro-Wilk normality test, the Wilcoxon signed-rank test was employed instead. The comparison test was applied across all models listed in the two tables to assess

the overall impact of the query methodology on model performance, rather than focusing on a specific model or type of model. The results in Table 5 show that the performance of the models in terms of MAP scores and nDGC has changed notably between the two scenarios described in the tables and not due to random chance. The significant differences underline the impact of query methodology on IR, particularly in specialised fields such as Spanish healthcare. This highlights the need for human-in-the-loop approaches where native speaker experts can embed their knowledge and experiences in how other authors use the language when writing papers.

Table 5: Statistical comparison between Table 3 and Table 4

Dataset	Test Type	Statistic	p-value
SB (MAP)	Paired t-test	14.24	7.05E-09
SB (nDCG)	Paired t-test	9.58	5.67E-07
SM (MAP)	Paired t-test	11.04	1.22E-07
SM (nDCG)	Paired t-test	9.18	8.90E-07
PO (MAP)	Wilcoxon	0	2.44E-04
PO (nDCG)	Wilcoxon	0	2.44E-04

6 Conclusion and Research Implication

The primary goal of this work was to present a foundational IR model for future research for biomedical Spanish systems, marking an initial step towards extending to other understudied languages. Additionally, the work sought to provide insights into the impact of query formation on such systems, potentially to guide the development of future query formation tools in such languages. Our research findings highlight the potential of hybrid models and re-rankers in enhancing the retrieval of biomedical literature in Spanish, thus proposing a potential benchmark for forthcoming Spanish IR methodologies. Furthermore, our research reveals a notable distinction in query formulation between English and Spanish. In English SR retrieval, where queries can be coined from the title/research question and used to retrieve relevant studies, we show that this is different in Spanish. In that, the hybrid models or future IR models for Spanish literature work may work well with the outcome of the study. As such, querying Spanish databases with synonyms of the outcome of the study may help in retrieving relevant information. Overall, this research contributes to the ongoing efforts to improve the comprehensiveness and precision of literature retrieval in SRs, particularly in LoE contexts.

7 Limitation of study and future works

This study focuses primarily on Spanish biomedical literature focusing on Mexican childhood health and specific database, LILACS. Thus, the specialisation, while valuable, limits the generalisability of the findings to LoE SRs. Future studies will explore the same methodology in different languages and databases to understand how these techniques generalise across various contexts and LoE. In addition, this study is a preliminary study to provide a foundational benchmark for future research and part of an ongoing Mexican project, COMO, thus to the best of our knowledge the first of its kind. As such, we worked with the available SR dataset provided by the COMO which is limited in size (8-8-15 queries for the three datasets). A future work will be to expand these queries and obtain more SR evaluation dataset. Also, resulting from the fact that this study is preliminary and foundational, we considered only four Spanish PLMs built on BERT and Roberta. However, other PLMs could offer different insights but are not explored in this study. Furthermore, this study compares two specific types of query formulations (original query vs. expert modifications-synonyms of outcome). This may not encompass the full range of possible query formulations, such as automated query expansion techniques. Future research could explore a wider array of query formulation strategies, including automated methods, to understand their impact on retrieval performance. Finally, a future work will be perform a parallel English IR approach on a known benchmark to help gauge the performance in a more apples-to-apples manner or using a cross-lingual supervision methods.

8 Acknowledgement

The authors would like to thank members of the Childhood Obesity in Mexico (COMO)[12] project for supporting this research.

References

Magaly Aceves-Martins, Naara L Godina-Flores, Yareni Yunuen Gutierrez-Gómez, Derek Richards,

[12]https://www.comoprojectmx.com/collaborators

Lizet López-Cruz, Marcela García-Botello, and Carlos Francisco Moreno-García. 2022. Obesity and oral health in mexican children and adolescents: systematic review and meta-analysis. *Nutrition Reviews*, 80(6):1694–1710.

Magaly Aceves-Martins, Lizet López-Cruz, Marcela García-Botello, Yareni Yunuen Gutierrez-Gómez, and Carlos Francisco Moreno-García. 2021. Interventions to prevent obesity in mexican children and adolescents: Systematic review. *Prevention Science*, 23(4):563–586.

Claudio Aracena and Jocelyn Dunstan. 2023. Development of pre-trained language models for clinical nlp in spanish. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Tanmoy Chakraborty, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2023. Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing*, 557:126680.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18. ACM.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Zhe Dong, Jianmo Ni, Dan Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. 2022. Exploring dual encoder architectures for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kate Flemming, Andrew Booth, Karin Hannes, Margaret Cargo, and Jane Noyes. 2018. Cochrane qualitative and implementation methods group guidance series—paper 6: reporting guidelines for qualitative, implementation, and process evaluation evidence syntheses. *Journal of Clinical Epidemiology*, 97:79–85.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. *Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline*, page 280–286. Springer International Publishing.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Naara L Godina-Flores, Yareni Yunuen Gutierrez-Gómez, Marcela García-Botello, Lizet López-Cruz, Carlos Francisco Moreno-García, and Magaly Aceves-Martins. 2022. Obesity and its association with mental health among mexican children and adolescents: systematic review. *Nutrition Reviews*, 81(6):658–669.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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)*. Association for Computational Linguistics.

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

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.

Guillermo Lopez-Garcia, Jose M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. Transformers for clinical coding in spanish. *IEEE Access*, 9:72387–72397.

Jing Lu, Keith Hall, Ji Ma, and Jianmo Ni. 2022a. Hyrr: Hybrid infused reranking for passage retrieval.

Jing Lu, Ji Ma, and Keith B. Hall. 2022b. Zero-shot hybrid retrieval and reranking models for biomedical literature. In *Conference and Labs of the Evaluation Forum*.

Ji Ma, Ivan Korotkov, Keith B. Hall, and Ryan T. McDonald. 2020a. Hybrid first-stage retrieval models for biomedical literature. In *Conference and Labs of the Evaluation Forum*.

Ji Ma, Ivan Korotkov, Keith B. Hall, and Ryan T. McDonald. 2020b. Hybrid first-stage retrieval models for biomedical literature. In *Conference and Labs of the Evaluation Forum*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Pedro Machado. 2016. Repositioning africa within the global. *Africa Today*, 63(2):88.

Gilberto Mercado-Mercado. 2023. Childhood obesity in mexico: A constant struggle and reflection for its prevention on the influence of family and social habits. *Obesity Medicine*, 44:100521.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search.

Lauge Neimann Rasmussen and Paul Montgomery. 2018. The prevalence of and factors associated with inclusion of non-english language studies in campbell systematic reviews: a survey and meta-epidemiological study. *Systematic Reviews*, 7(1).

Sameh Neji, Tarek Chenaina, Abdullah M. Shoeb, and Leila Ben Ayed. 2021. Hyra: An effective hybrid ranking model. *Procedia Computer Science*, 192:1111–1120.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert.

Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1–22.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

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

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yulia Shenderovich, Manuel Eisner, Christopher Mikton, Frances Gardner, Jianghong Liu, and Joseph Murray. 2016. Methods for conducting systematic reviews of risk factors in low- and middle-income countries. *BMC Medical Research Methodology*, 16(1).

Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Chantelle Garritty, Tamara Rader, and David Moher. 2010. Updating systematic reviews.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Sarah Catherine Walpole. 2019. Including papers in languages other than english in systematic reviews: important, feasible, yet often omitted. *Journal of Clinical Epidemiology*, 111:127–134.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Shucong Zhang, Malcolm Chadwick, Alberto Ramos, and Sourav Bhattacharya. 2022. Cross-attention is all you need: Real-time streaming transformers for personalised speech enhancement.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings*

of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23. ACM.