Hybrid IR Pipeline (Retrieve and Rerank) for PRISMA Screening Automation in Systematic Literature Reviews

José Jesús Alejandro Soria Vásquez[1]; Maximiliano Rodolfo Arroyo Ulloa[2]; Diana Peche Cieza[3]
[1,2,3] Universidad Católica Santo Toribio de Mogrovejo (USAT), Chiclayo, Peru, 72624280@usat.pe, marroyo@usat.pe, dpeche@usat.pe

Abstract– ***The purpose of this research was to design, implement, and validate a hybrid Information Retrieval (IR) pipeline based on a Retrieve-and-Rerank architecture to automate the PRISMA screening phase in systematic literature reviews (SLR). Manual screening consumes on average 163 researcher-hours for corpora on the order of 10,000 records, and a single SLR demands 67.3 weeks on average with an average inclusion rate of only 2.94%. To address this problem, three complementary stages were chained: BM25 Okapi for lexical retrieval, SBERT all-mpnet-base-v2 for dense semantic retrieval, and a Cross-Encoder ms-marco-MiniLM-L-6-v2 for neural re-ranking. The system was implemented in 748 lines of Python on Google Colab (GPU T4) under an open-source, fully reproducible, zero-shot framework, and validated against the Gold Standard of Liu et al. (2023), an SLR on Vehicle Routing Problems with Time Windows (VRPTW) with 64 retrievable articles from a clean corpus of 4,712 unique records. Results showed a Precision@10 of 40.0%, an adjusted Recall of 54.7% (35/64 articles), and a Mean Reciprocal Rank of 1.0000. An ablation study over four experimental configurations confirmed the irreducibility of each component: removing BM25 reduced Recall by 37.9%, removing SBERT by 13.5%, and removing the Cross-Encoder collapsed NDCG@10 from 0.4479 to 0.1799. Finally, the pipeline autonomously discarded 82.9% of the corpus in under 3 minutes (2.4 minutes processing time on GPU), reducing total screening effort between an estimated 91.8% (13.4 hours) and 98.0% (3.3 hours) compared to the conventional manual method, and formal compliance with the PRISMA-trAIce guideline (2025) was fully documented.***

Keywords-- ***Systematic literature review, PRISMA-trAIce, information retrieval, Retrieve and Rerank, sentence transformers.***

Pipeline Híbrido de Recuperación de Información (Retrieve and Rerank) para Automatizar el Screening PRISMA en Revisiones Sistemáticas

José Jesús Alejandro Soria Vásquez[1], Maximiliano Rodolfo Arroyo Ulloa, Doctor[2],Diana Peche Cieza, Magister[3]
[1,2,3] Universidad Católica Santo Toribio de Mogrovejo (USAT), Chiclayo, Perú, 72624280@usat.pe, marroyo@usat.pe, dpeche@usat.pe

Resumen– ***El propósito de esta investigación fue diseñar, implementar y validar un pipeline híbrido de Recuperación de Información (IR) basado en una arquitectura Retrieve-and-Rerank para automatizar la fase de screening PRISMA en revisiones sistemáticas de literatura (SLR). El cribado manual consume en promedio 163 horas-investigador para corpus del orden de 10 000 registros, y una SLR demanda 67,3 semanas en promedio con una tasa media de inclusión de apenas 2,94 %. Para abordar esta problemática, se encadenaron tres etapas complementarias: BM25 Okapi para recuperación léxica, SBERT all-mpnet-base-v2 para recuperación semántica densa, y un Cross-Encoder ms-marco-MiniLM-L-6-v2 para re-ranking neuronal. El sistema se implementó en 748 líneas de Python sobre Google Colab (GPU T4) bajo una estrategia zero-shot de costo cero, y se validó contra el Gold Standard de Liu et al. (2023), una SLR sobre Vehicle Routing Problems with Time Windows (VRPTW) con 64 artículos recuperables de un corpus de 4 712 registros. Los resultados mostraron una Precision@10 de 40,0 %, Recall ajustado de 54,7 % (35/64 artículos) y un MRR de 1,0000. La ablación sobre cuatro configuraciones confirmó la irreductibilidad de cada componente: eliminar BM25 redujo el Recall un 37,9 %, eliminar SBERT un 13,5 %, y eliminar el Cross-Encoder colapsó el NDCG@10 de 0,4479 a 0,1799. Finalmente, el pipeline descartó autónomamente el 82,9 % del corpus en menos de 3 minutos, reduciendo el esfuerzo de screening entre un estimado de 91,8 % y 98,0 % respecto al método manual convencional, y se documentó el cumplimiento de la guía PRISMA-trAIce (2025).***

Palabras clave-- Revisión sistemática de literatura, PRISMA-trAIce, recuperación de información, Retrieve and Rerank, sentence transformers.

I. Introducción

La producción científica global desbordó la capacidad de los métodos tradicionales de síntesis de evidencia [1]. En las revisiones sistemáticas de literatura (SLR), la fase de screening se convirtió en el cuello de botella más costoso del proceso PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [2]. Un investigador enfrenta corpus iniciales de miles de registros tras la búsqueda booleana en Scopus, Web of Science y EBSCO [3]; una SLR demanda en promedio 67,3 semanas, con casos extremos de 186 semanas [4], y la tasa media de inclusión es solo del 2,94 % de los estudios identificados [4]. Esto provoca obsolescencia, errores por fatiga cognitiva [5] y exclusión de equipos con recursos limitados [6].

En América Latina, la brecha tecnológica amplifica esta crisis: mientras el hemisferio norte dispone de Covidence, Rayyan o Elicit, los grupos locales se confinan al método booleano, que adolece de vocabulary mismatch: si el autor usó un sinónimo no previsto, el registro se pierde [7]. La falta de exhaustividad introduce sesgos y compromete la validez [8]. Los modelos Transformer como SBERT y Cross-Encoders permiten evaluar relevancia sin etiquetado manual (zero-shot) [9].

Sin embargo, la literatura presenta un vacío: ASReview emplea active learning que requiere etiquetado inicial [10]; Rayyan ofrece clasificación binaria iterativa [11]; y los LLMs generativos (como PROMPTHEUS [12]) dependen de APIs comerciales. Ninguna solución combina las tres capas de recuperación (léxica, semántica y neuronal) en un pipeline unificado zero-shot, sin costo y con transparencia total.

La pregunta de investigación es: ¿en qué medida un pipeline híbrido IR de tres etapas (BM25 + SBERT + Cross-Encoder) automatiza el screening PRISMA reduciendo el esfuerzo humano sin comprometer la recuperación? Los objetivos específicos son: (a) estructurar un corpus replicando la estrategia de [13]; (b) implementar la arquitectura híbrida sobre infraestructura de costo cero (Google Colab, GPU T4); y (c) evaluar el rendimiento mediante métricas IR (Recall@K, Precision@K, MRR, NDCG) y un estudio de ablación. Las contribuciones son: una arquitectura Sparse-Dense-Rerank validada contra un Gold Standard de 64 artículos; un estudio de ablación que demuestra la irreductibilidad de cada componente; el cumplimiento documentado de PRISMA-trAIce [14]; y una herramienta de costo cero y código abierto que democratiza el acceso a tecnología de screening.

II. Revisión de Literatura

A. Evolución de las herramientas de screening automatizado

La automatización del screening ha transitado por cuatro paradigmas. El booleano, nativo en Scopus y PubMed, garantiza precisión léxica pero adolece de vocabulary mismatch, fenómeno donde un artículo relevante se pierde si los autores usan sinónimos distintos a las palabras clave de la búsqueda [7]. El supervisado, introducido por el estudio de [15] con Vectores de Soporte, logró 40–50 % de reducción pero

requiere datos etiquetados. El active learning, exponente ASReview [10], aprende iterativamente del revisor; en [16] se implementó Research Screener, aunque dependiente de "artículos semilla" iniciales. El Transformer zero-shot, aplicado por la investigación de [17] con BART, eliminó la anotación pero operó en una sola etapa sin recuperación previa. El paradigma híbrido Sparse-Dense-Rerank [18] integra múltiples capas: en [19] se usaron modelos densos BERT para capturar similitudes conceptuales abstractas; el estudio de [12] presentó PROMPTHEUS con LLMs generativos; En [20] se propuso dense retrieval con retroalimentación es decir que calibra los vectores dinámicamente mediante retroalimentación continua del usuario; y [21] confirmó que la recuperación híbrida es el estado del arte emergente.

B. Justificación teórica de los modelos seleccionados para dominios transversales

Un dominio transversal es aquel que no se restringe a una disciplina particular; en este estudio implica que el pipeline debe ser aplicable a SLR de ingeniería, medicina o ciencias sociales sin reentrenamiento. La idoneidad de all-mpnet-base-v2 [22] y ms-marco-MiniLM-L-6-v2 [23] se fundamenta en tres propiedades. Primero, MPNet combina masked y permuted language modeling, mitigando el sesgo posicional de BERT y produciendo representaciones robustas frente a variabilidad terminológica en dominios no vistos [22]; se pre-entrenó sobre más de 160 GB de texto y, en all-mpnet-base-v2, se afina sobre mil millones de pares de oraciones mediante contrastive learning, generando embeddings de 768 dimensiones bajo la arquitectura bi-encoder de [24]. Segundo, el Cross-Encoder ms-marco-MiniLM-L-6-v2 se obtiene mediante Deep Self-Attention Distillation, técnica task-agnostic que comprime un modelo docente en uno de 6 capas y 23 millones de parámetros preservando interdependencias token-a-token [23]; fue afinado sobre MS-MARCO Passage Ranking (más de 500 000 consultas de Bing y 8,8 millones de pasajes), transfiriendo capacidad de re-ranking cross-domain. Tercero, la arquitectura Cross-Encoder supera a los bi-encoders al procesar consulta y documento como par concatenado [25]. Estas propiedades habilitan la transferencia directa a cualquier dominio sin reentrenamiento, condición necesaria para su operación zero-shot.

C. Identificación de la brecha en la literatura

Pese a estos avances, ningún trabajo combina recuperación léxica (BM25), semántica densa (SBERT) y re-ranking neuronal (Cross-Encoder) en una arquitectura zero-shot, de costo cero y con trazabilidad PRISMA. La Ref. [26] presentó un sistema monolingüe de doble etapa que depende de refinamientos manuales; y en [27] se presentó EmbedSLR, basado en embeddings pero sin re-ranking neuronal.

III. MATERIALES Y MÉTODOS

La presente investigación es de tipo aplicada, dado que su objetivo no es generar teoría, sino desarrollar y validar una herramienta computacional que resuelve un problema práctico: la automatización del screening PRISMA en revisiones sistemáticas de literatura. El enfoque es cuantitativo, pues todas las conclusiones se sustentan en métricas numéricas estándar de Information Retrieval (Recall@K, Precision@K, MRR, NDCG@K, tiempo de ejecución y porcentaje de reducción del esfuerzo humano). El diseño es experimental computacional, ya que se implementó un algoritmo híbrido Retrieve-and-Rerank que fue ejecutado bajo cuatro configuraciones distintas (estudio de ablación C1–C4) para comparar objetivamente sus resultados. La validación se realizó mediante Gold Standard, contrastando el desempeño del pipeline contra la SLR de [13], considerada como verdad de referencia para medir la recuperación de artículos relevantes. La unidad de análisis fue el corpus documental de 4 712 artículos, y la muestra comprendió el pool candidato de 807 documentos evaluado contra un Gold Standard de 64 artículos.

A. Visión general

El pipeline opera bajo el patrón arquitectónico Sparse-Dense-Rerank [18]. El flujo se ejecuta en tres fases secuenciales sobre un corpus previamente sometido a un proceso ETL (Extract-Transform-Load) de deduplicación y normalización. La entrada consiste en: (a) un corpus de metadatos bibliográficos exportados de Scopus, Web of Science y EBSCO (título, abstract, autores, año, DOI, palabras clave); y (b) un thesis target, párrafo extenso de ingeniería de prompts que codifica la pregunta de investigación, variables de interés, criterios de inclusión y cláusulas de exclusión explícitas.

A diferencia de una cadena booleana, el thesis target contiene entre 200 y 400 palabras que describen con precisión técnica el alcance de la revisión, incluyendo instrucciones negativas (negative prompting). Su estructura traduce los objetivos y criterios de exclusión del Gold Standard [13] a un prompt técnico. Al inyectar la taxonomía estándar del estado del arte en ruteo de vehículos, el sistema establece anclas semánticas precisas para evaluar su capacidad de recuperación de forma autónoma (zero-shot). El thesis target se compone de cuatro segmentos funcionales, cuya disección se presenta en la Tabla I.

B. Fase 1: Recuperación Léxica (BM25 Okapi)

La primera fase emplea el algoritmo BM25 Okapi [28], modelo probabilístico que cuantifica la relevancia de un documento respecto a una consulta mediante coincidencia exacta de términos. La puntuación se calcula como:

$$\text{Score} = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)} \quad (1)$$

Donde $f(q_i,D)$ es la frecuencia del término q_i en el documento D, $|D|$ es la longitud del documento, y avgdl es la longitud promedio del corpus. Los parámetros $k_1 = 1{,}5$ y $b = 0{,}75$ son los valores estándar ampliamente adoptados por la comunidad de Recuperación de Información, derivados de la

experimentación empírica sobre los benchmarks TREC que consolidó el framework probabilístico de [28]; b = 0,75 constituye el consenso de facto para la normalización por longitud, mientras que k_1 = 1,5 se ubica dentro del rango operacional recomendado [1,2; 2,0]. El factor IDF(q_i) pondera la importancia inversa de cada término en el corpus, penalizando palabras de alta frecuencia y amplificando las discriminantes.

TABLA I
DISECCIÓN DEL THESIS TARGET

Segmento	Contenido	Función en el pipeline	Origen en [13]
Dominio y objetivo	identify, evaluate, and compare algorithmic solutions for VRPTW	Centra el vector SBERT en optimización combinatoria, descartando gestión administrativa genérica	Sección 2.1 (Research Questions)
Taxonomía de algoritmos	ALNS, Genetic Algorithms, Simulated Annealing, Tabu Search, ACO	Almacena "oro semántico": el Cross-Encoder bonifica artículos que mencionan estas siglas	Clasificación teórica estándar VRPTW
Variantes logísticas	capacitated vehicles, heterogeneous fleets, dynamic demand, green routing	Prioriza formulaciones modernas sobre trabajos puramente teóricos	Alcance del problema (Introducción GS)
Penalidad negativa	Articles lacking algorithmic validation... are strictly excluded	Activa el castigo logarítmico del Cross-Encoder: artículos sin validación reciben logits negativos	Tabla 2 (Exclusion Criteria #1 y #2)

La función de BM25 es recuperar siglas técnicas exactas ("VRPTW", "MTO", "EOQ") que los modelos semánticos pueden diluir. El resultado es un conjunto ordenado S_BM25 (K = 500) sobre texto plano unificado (título + resumen), descartando BM25F [28] por redundancia en esta fase.

C. Fase 2: Recuperación Semántica Densa (SBERT)

La segunda fase utiliza la arquitectura Sentence-BERT [24] implementada sobre el modelo all-mpnet-base-v2 [22], Transformer pre-entrenado de 110 millones de parámetros que proyecta textos a un espacio vectorial de 768 dimensiones. Cada documento del corpus y el thesis target se codifican como vectores densos, y la relevancia se cuantifica mediante similitud coseno:

$$\cos(\theta) = \frac{\sum_{i=1}^{768} A_i B_i}{\sqrt{\sum_{i=1}^{768} A_i^2} \cdot \sqrt{\sum_{i=1}^{768} B_i^2}} \quad (2)$$

Donde A es el vector del thesis target y B el de cada documento. El rango es [−1, 1], donde 1 indica identidad semántica y 0 ortogonalidad. La función de SBERT es resolver el vocabulary mismatch: un artículo que use "fleet scheduling" en lugar de "vehicle routing" recibirá puntuación BM25 baja pero similitud coseno alta. El resultado es un conjunto ordenado S_SBERT (K = 500). La unión (S_BM25 ∪ S_SBERT) generó 807 documentos únicos a partir de dos conjuntos de 500, con 193 capturados por ambos filtros.

D. Fase 3: Re-Ranking Neuronal (Cross-Encoder)

La tercera fase aplica el Cross-Encoder ms-marco-MiniLM-L-6-v2 [23] sobre el pool candidato. A diferencia de SBERT (bi-encoder) [25], procesa consulta y documento como un par concatenado, capturando interdependencias token-a-token entre la pregunta y cada abstract.

El Cross-Encoder recibe el par [thesis_target, corpus_text$_i$] y produce un logit crudo $z_i \in \mathbb{R}$, típicamente en el rango [−10, +10], que representa la relevancia absoluta del documento. Este logit se transforma a escala probabilística interpretable mediante la función Sigmoid estándar:

$$\text{Rel. Abs.}(\%) = \sigma(z) \times 100 = \frac{1}{1+e^{-z}} \times 100 \quad (3)$$

Esta transformación mapea los logits a [0%, 100%]; un artículo con z = 1,83 obtiene 86,2% de Relevancia Absoluta. Adicionalmente, se calcula un Match Score relativo mediante normalización Min-Max:

$$\text{Match_Score}(\%) = \frac{z_i - z_{min}}{z_{max} - z_{min}} \times 100 \quad (4)$$

El Match Score indica la posición relativa en la corrida, mientras que la Relevancia Absoluta ofrece una interpretación probabilística independiente del pool.

E. Proceso ETL y preparación del corpus

Antes de la ejecución, los datos se someten a un proceso ETL automatizado: (a) Extracción de archivos heterogéneos (.csv de Scopus, .xls de WoS, .txt de EBSCO) con detección automática de codificación; (b) Transformación a un esquema unificado (title, abstract, authors, year, doi, all_keywords, source_file) y concatenación del campo corpus_text = título + ". " + abstract; y (c) Carga, deduplicación que eliminó 2 149 duplicados del corpus original de 6 861 registros, produciendo 4 712 artículos únicos. La composición se detalla en la Tabla II.

TABLA II
COMPOSICIÓN DEL CORPUS

Fuente de Exportación	Formato	Registros crudos
Scopus	.csv	4 229
Web of Science - Core Collection (Parte 1)	.xls	1 000
Web of Science - Core Collection (Parte 2)	.xls	204
Web of Science - All Databases (Parte 1)	.xls	998
Web of Science - All Databases (Parte 2)	.xls	430
Total crudo		6 861
Duplicados eliminados (ETL)		−2 149
Corpus limpio (N)		4 712

F. Fusión por unión simple

La fusión de S_BM25 y S_SBERT se realiza mediante unión simple con etiquetado del canal de captura ("Solo BM25", "Solo SBERT" o "AMBOS"). Se descartó el método

Reciprocal Rank Fusion (RRF) porque penaliza matemáticamente las "joyas asimétricas" (artículos que solo un canal detecta): un paper altamente relevante ausente en el segundo ranking recibe un score menor que un paper mediocre presente en ambos. En una arquitectura con Cross-Encoder al final, RRF resulta contraproducente: el Cross-Encoder es el juez supremo de relevancia y no requiere que una fórmula de consenso filtre candidatos antes de su evaluación. La unión simple garantiza que el Top-K completo de ambos motores pase intacto al re-ranking neuronal.

G. Estrategia de búsqueda y recolección del corpus

La recolección primaria del corpus replicó fielmente la cadena booleana oficial de [13], introducida en Scopus y Web of Science: ("vehicle routing problems" OR "VRP" OR "vehicle routing problems with time windows" OR "VRPTW" OR "vehicle routing problems with time window constraints" OR "VRP with time constraints" OR "VRP with time windows" OR "VRP with time constraints" OR "VRP with time window constraints") Los filtros de refinamiento se aplicaron según la Tabla III.

TABLA III
CRITERIOS DE FILTRO APLICADOS

Parámetro	Valor aplicado
Frontera temporal	2018–2022
Tipo de documento (WoS)	Article, Proceeding Paper
Tipo de documento (Scopus)	Article, Conference Paper
Tipos excluidos	Review, Book chapter, Conference review, Note, Letter, Short survey, Erratum, Editorial Material, Early Access
Idioma	English
Tipo de fuente	Sin filtro (Journals + Conference Proceedings)

Fuente: [13]

La aplicación de esta matriz sobre ambas bases produjo 6 861 registros brutos antes de la deduplicación. Para validar el pipeline se adoptó una estrategia de validación retrospectiva contra la SLR de [13], publicada en Sustainability en 2023, que revisa la literatura sobre VRPTW con metodología PRISMA convencional. La auditoría de su bibliografía (77 referencias) determinó la población real analizable, presentada en la Tabla IV.

TABLA IV
AUDITORÍA DEL GOLD STANDARD

Categoría	Cantidad
Total de referencias en bibliografía de [13]	77
Exclusión: guías metodológicas y trabajos futuros	−10
Total del bloque SLR reportado	67
Exclusión: citas solo en la Introducción	−2
Artículos reales analizados en el Gold Standard	65
Artículo ausente del corpus (no ingerido en datasets)	−1
Artículos recuperables (input físico en la base de datos)	64

Las exclusiones se justifican por razones metodológicas formales: las referencias internas n.º 59 y n.º 60 del artículo original de [13] son manuales metodológicos (guía PRISMA y protocolo Kitchenham), no literatura primaria sobre VRPTW; las referencias internas n.º 70 a n.º 77 corresponden a citas de tendencias futuras fuera de la muestra de revisión; y las referencias internas n.º 1 y n.º 11 se mencionan exclusivamente en la contextualización introductoria, sin participar en el análisis sistemático.

H. Configuraciones del estudio de ablación

En alineación con el tercer objetivo específico, se ejecutaron cuatro configuraciones experimentales (Tabla V) para demostrar que cada componente del pipeline es necesario y no redundante.

TABLA V
CONFIGURACIONES DE ABLACIÓN

ID	Configuración	Fases activas	Pool candidato
C1	Pipeline Completo	BM25 + SBERT + Cross-Encoder	807
C2	Sin BM25	SBERT + Cross-Encoder	500
C3	Sin SBERT	BM25 + Cross-Encoder	500
C4	Sin Cross-Encoder	BM25 + SBERT (sin re-ranking)	807

En C1, la unión de los Top-500 de BM25 y SBERT generó 807 documentos únicos. En C2, solo se evaluaron los 500 documentos recuperados por SBERT. En C3, solo los 500 de BM25. En C4, se conservó el pool de 807 pero sin aplicar el Cross-Encoder, utilizando en su lugar un promedio ponderado de los scores normalizados de BM25 y SBERT como criterio de ranking.

I. Métricas de evaluación

Se emplearon cuatro métricas estándar de Information Retrieval:

Precision@K: fracción de documentos del Gold Standard entre los primeros K del ranking.

$$\mathrm{P@K} = \frac{|\mathrm{GS} \cap \mathrm{Top_K}|}{\mathrm{K}} \quad (5)$$

Recall@K: fracción del Gold Standard recuperada en los primeros K posiciones. , donde $|\ \mathrm{GS}\ | = 64$.

$$\mathrm{R@K} = \frac{|\mathrm{GS} \cap \mathrm{Top_K}|}{|\mathrm{GS}|} (6)$$

MRR (Mean Reciprocal Rank): inverso de la posición del primer documento del Gold Standard en el ranking.

$$\mathrm{MRR} = \frac{1}{\mathrm{rank_{primer_GS}}} (7)$$

NDCG@K (Normalized Discounted Cumulative Gain): mide la calidad del ranking ponderando la posición de los documentos relevantes con un factor de descuento logarítmico. Se define como $\mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}$, donde la Ganancia Acumulada (DCG) es:

$$\mathrm{DCG@K} = \sum_{i=1}^{K} \frac{\mathrm{rel_i}}{\log_2(i+1)} \quad (8)$$

El diagrama de flujo de la Fig. 1 documenta la arquitectura de recuperación y las decisiones de exclusión del pipeline IR, integrando las tres fases secuenciales (BM25, SBERT, Cross-Encoder) sobre el corpus ETL y el thesis target, con el pool candidato final y la matriz de salida

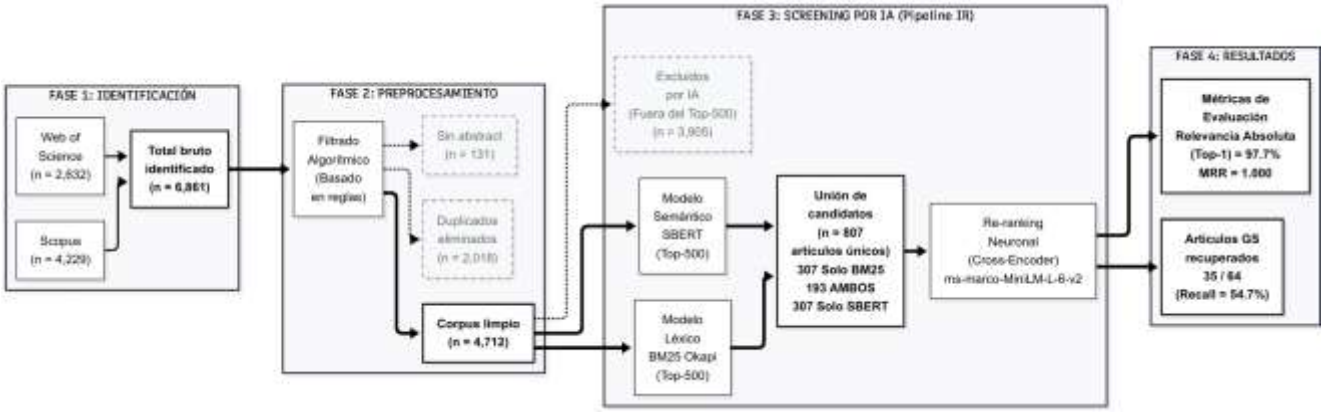


Fig. 1 Diagrama de flujo arquitectónico del pipeline IR

IV. Resultados

A. Rendimiento global del pipeline (Configuración C1)

La ejecución del pipeline completo (BM25 + SBERT + Cross-Encoder) sobre el corpus de 4 712 artículos con K = 500 por fase produjo un pool candidato de 807 documentos únicos. La distribución por canal de captura fue: 307 artículos recuperados exclusivamente por BM25 (38,0 %), 307 exclusivamente por SBERT (38,0 %) y 193 capturados por ambos filtros simultáneamente (23,9 %). Esta distribución simétrica confirma la complementariedad de ambos mecanismos, dado que cada uno aporta un volumen equivalente de candidatos que el otro no habría detectado. El Cross-Encoder evaluó los 807 candidatos y generó logits en el rango [−3,87; +3,76], que tras la transformación Sigmoid produjeron Relevancias Absolutas entre 2,1 % y 97,7 %. El tiempo total de procesamiento fue de 2,4 minutos sobre GPU T4, incluyendo las tres fases y la generación de la matriz Excel de salida.

TABLA VI
MÉTRICAS DE INFORMATION RETRIEVAL, PIPELINE COMPLETO (C1)

K (Top Rank)	Precision@K	Recall@K	GS encontrados
10	40,0%	6,2%	4 / 64
20	20,0%	6,2%	4 / 64
50	12,0%	9,4%	6 / 64
100	12,0%	18,8%	12 / 64
200	9,5%	29,7%	19 / 64
807 (pool completo)	4,3%	54,7%	35 / 64

El valor de Precision@10 = 40,0 % indica que 4 de cada 10 artículos en la cúspide del ranking pertenecen al Gold Standard. El MRR alcanzó el valor máximo de 1, lo que significa que el primer artículo del ranking es un documento del Gold Standard (referencia interna de [13]: "A hyper-heuristic with two guidance indicators for bi-objective mixed-shift vehicle routing problem with time windows", con Relevancia Absoluta del 97,7 %).

B. Resultados del estudio de ablación

TABLA VII
ESTUDIO DE ABLACIÓN. COMPARATIVA COMPLETA DE CONFIGURACIONES

Config	Pool	GS	P@10	P@20	P@50	R@10	R@20	R@50	R@100	R@200	R@Full	MRR	NDCG@10	NDCG@50
C1 Completo	807	37	40,0%	20,0%	16,0%	6,2%	6,2%	12,5%	21,9%	32,8%	57,8%	1	0,4479	0,2689
C2 Sin BM25	500	23	30,0%	20,0%	18,0%	4,7%	6,2%	14,1%	21,9%	29,7%	35,9%	1	0,3786	0,3844
C3 Sin SBERT	500	32	40,0%	25,0%	24,0%	6,2%	7,8%	18,8%	25,0%	31,2%	50,0%	1	0,4819	0,3975
C4 Sin CE	807	37	20,0%	10,0%	8,0%	3,1%	3,1%	6,2%	14,1%	23,4%	57,8%	0,25	0,1799	0,1140

Fuente: Elaboración propia

Nota: GS reporta el conteo algorítmico bruto (37) mediante substring matching imparcial. El Recall ajustado humanamente depurado es 35/64 = 54.7%.

La Tabla VII presenta la comparativa completa. Los resultados revelaron tres hallazgos. Hallazgo 1: BM25 es indispensable para maximizar el Recall. La eliminación de BM25 (C2) provocó la caída más severa en Recall, de 57,8 % a 35,9 %, una reducción del 37,9 %, debido a que BM25 captura artículos con terminología exacta ("VRPTW", "time windows",

"metaheuristic") que SBERT puede diluir en el espacio semántico. De los 35 artículos del Gold Standard depurados capturados por el pipeline completo, 14 fueron recuperados exclusivamente por BM25. Hallazgo 2: SBERT aporta cobertura semántica insustituible. Sin SBERT (C3), el Recall descendió a 50,0 %, una reducción del 13,5 %. Aunque BM25 es individualmente más potente para este dominio técnico, SBERT recupera artículos que utilizan terminología alternativa ("fleet scheduling", "logistics dispatching", "route optimization"), sinónimos conceptuales que BM25 no puede detectar por diseño. En consecuencia, 5 artículos del Gold Standard fueron recuperados exclusivamente por SBERT. Hallazgo 3: el Cross-Encoder es el diferenciador de calidad del ranking. Sin el Cross-Encoder (C4), el pool candidato permanece idéntico (807 documentos, 37 GS algorítmicos), pero el NDCG@10 colapsa de 0,4479 a 0,1799, una reducción del 59,8 %. El MRR cae de 1 a 0,25 , lo que indica que sin re-ranking neuronal, el primer artículo del Gold Standard aparece en la posición 4 en lugar de la posición 1.

Paradoja del NDCG: Se denomina paradoja del NDCG al fenómeno contraintuitivo mediante el cual una configuración arquitectónicamente inferior produce un valor de NDCG@10 aparentemente superior. En este estudio, la configuración C3 (Sin SBERT) presenta NDCG@10 = 0,4819 frente a 0,4479 del pipeline completo C1. Esto no indica superioridad de C3, sino un trade-off matemático inherente: al eliminar SBERT, el pool se reduce a 500 artículos exclusivamente léxicos, facilitando que los Gold Standard (que contienen las siglas exactas) asciendan en el ranking sin competencia semántica. Sin embargo, este pool "limpio" sacrifica 5 artículos GS que solo SBERT detecta (Recall cae de 57,8 % a 50,0 %). Los artículos que SBERT introduce y que desplazan a los GS en las posiciones superiores son evaluados por la métrica NDCG como falsos positivos; no obstante, una auditoría técnica revela que son Heavy-Hitters semánticos: publicaciones que el Cross-Encoder puntúa con más del 90 % de probabilidad de relevancia y que abordan el mismo problema matemático bajo terminología alternativa, lo que sugiere que podrían corresponder a posibles omisiones en la SLR humana original.

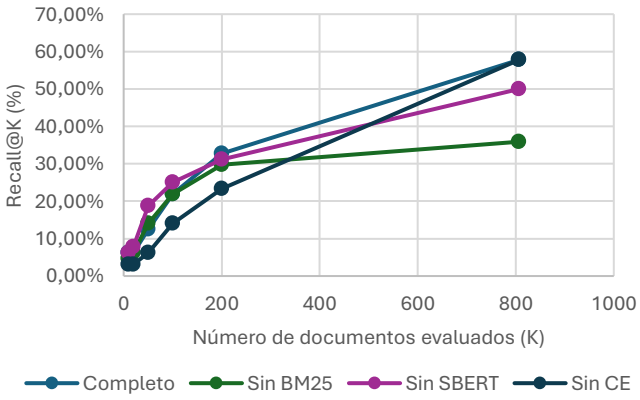


Fig. 2 Curva Recall@K del estudio de ablación

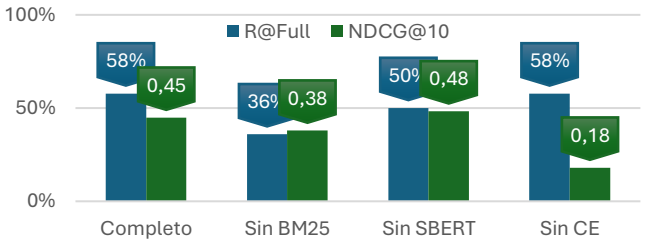


Fig. 3 Métricas de recuperación (R@Full) y ordenamiento (NDCG@10)

C. Distribución por canal de captura

TABLA VIII
DISTRIBUCIÓN DE ARTÍCULOS POR CANAL DE CAPTURA

Canal	Artículos en pool (807)	Artículos GS capturados (35)
Solo BM25 (Léxico)	307 (38,0%)	14 (40,0%)
Solo SBERT (Semántico)	307 (38,0%)	5 (14,3%)
AMBOS filtros	193 (23,9%)	16 (45,7%)

La categoría "AMBOS" concentra el 45,7 % de los artículos del Gold Standard capturados, a pesar de representar solo el 23,9 % del pool total. Este dato confirma que los artículos más relevantes tienden a poseer tanto terminología exacta (alta puntuación BM25) como coherencia semántica profunda (alta similitud coseno), validando la hipótesis de que la intersección de ambos filtros es un predictor fuerte de relevancia.

D. Artículos del Gold Standard recuperados

La Tabla IX presenta la lista de los 10 artículos del Gold Standard mejor posicionados en el ranking del pipeline. Tras una depuración humana manual (verificación de coincidencia exacta de títulos), el conteo ajustado es de 35 artículos recuperados de 64 posibles (Recall = 54,7 %). La discrepancia entre 37 (algorítmico) y 35 (depurado) se debe a 2 coincidencias parciales por substring matching que la inspección manual descartó como falsos positivos.

Se destaca que el pipeline también recuperó 2 artículos de la sección de "Trabajos Futuros" de [13] (referencias internas n.º 71 y n.º 72 de su bibliografía), ubicándolos en posiciones 258 y 735 respectivamente. Estos fueron excluidos del cómputo de Recall por no pertenecer formalmente a la muestra de la SLR, pero su recuperación evidencia la capacidad del sistema para identificar literatura periférica relevante.

E. Análisis de sensibilidad del parámetro K

El valor K = 500 no fue seleccionado arbitrariamente. Se realizó una evaluación preliminar con K = 100 que produjo resultados severamente degradados: el pool candidato se redujo a 193 artículos únicos (frente a 807 con K = 500), el Recall descendió a valores por debajo del 25 % y múltiples artículos del Gold Standard que contenían formulaciones semánticas no convencionales fueron excluidos antes de alcanzar el Cross-Encoder. Este resultado confirma que K = 100 es insuficiente para un corpus de esta magnitud, dado que limita

excesivamente la cobertura de ambos motores de recuperación. El valor K = 500 representa un punto de operación empírico donde la cobertura de recuperación (807 documentos, 17,1 % del corpus) es suficiente para capturar la mayoría de los artículos relevantes sin saturar la capacidad computacional del Cross-Encoder en una sesión de Colab con GPU T4 (15 GB VRAM).

TABLA IX
TOP 10 ARTÍCULOS DEL GOLD STANDARD RECUPERADOS

#	Rank	Rel. Abs.	Canal	Año	Título (abreviado)
1	1	97,7%	AMBOS	2018	A hyper-heuristic for bi-objective MS-VRPTW
2	4	96,2%	BM25	2022	Data-Driven Bus Route Optimization (ALNS)
3	8	94,9%	AMBOS	2020	A hybrid swarm intelligence algorithm for VRPTW
4	10	94,7%	AMBOS	2020	Lexicographic two-stage algorithm for VRPSPDTW
5	21	92,6%	AMBOS	2018	Evolutionary scatter search PSO for VRPTW
6	29	92,0%	AMBOS	2018	Evolutionary scatter search PSO for VRPTW (var.)
7	40	90,5%	AMBOS	2020	Hybrid swarm intelligence for VRPTW (var.)
8	45	90,0%	AMBOS	2020	Heterogeneous VRP with TW and limited resources
9	52	89,5%	BM25	2021	Decomposition Multi-Objective VND for Logistics
10	55	89,1%	AMBOS	2019	Quality of service objectives for VRPTW

Nota: artículos #6 y #7 son variantes de capitalización del mismo registro.

Un incremento de K eleva el techo de Recall potencial, pero la curva Recall@K (Fig. 2) evidencia un comportamiento asintótico de rendimientos decrecientes: a partir de K = 500, las ganancias marginales son cada vez menores, indicando que la mayoría de los artículos del Gold Standard accesibles mediante recuperación léxica y semántica ya han sido capturados. En el límite teórico, configurar K igual al corpus total garantizaría un Recall del 100 %, pero anularía la función primordial del pipeline, dado que las fases de recuperación pierden su razón de ser como mecanismos de filtrado si todo el corpus se transfiere al Cross-Encoder. La selección de K responde, por tanto, a un trade-off entre cobertura, costo computacional y preservación del propósito funcional del pipeline como filtro de Identificación Primaria.

V. DISCUSIONES

A. Comparativa con herramientas existentes

El pipeline propuesto se diferencia de las herramientas existentes por combinar tres capas de recuperación (léxica, semántica y neuronal) en una arquitectura zero-shot, con costo cero y código abierto MIT. ASReview [10] emplea active learning que requiere etiquetado inicial; Rayyan [11] ofrece clasificación binaria iterativa; Elicit [29] automatiza la extracción de datos vía LLM; Consensus [30] genera respuestas sintetizadas a partir de papers; SciSpace [31] ofrece agentes de IA para navegación literaria; Covidence [32] y DistillerSR [33] son propietarios y de suscripción. Ninguna combina las tres capas en modo zero-shot con trazabilidad PRISMA-trAIce.

Se excluyen de la comparativa las herramientas de descubrimiento (Connected Papers, Research Rabbit, Litmaps) dado que operan como complementos de exploración y no como sistemas de screening PRISMA. No obstante, Semantic Scholar merece mención especial por ofrecer embeddings pre-entrenados (SPECTER2) que podrían sustituir a SBERT en futuras versiones del pipeline, siempre que el corpus se obtenga a través de su API en lugar de exportaciones offline.

La ventaja competitiva del pipeline reside en cuatro aspectos: (a) la ausencia total de entrenamiento previo (zero-shot), que en principio permite aplicarlo a diversos dominios sin datos de referencia; (b) la combinación de tres capas de recuperación que el estudio de ablación demostró como irreductible; (c) el costo monetario cero; y (d) la capacidad de procesar CSVs exportados de cualquier base de datos académica sin depender de APIs externas.

B. La paradoja del Recall

Un Recall de 54,7 % puede interpretarse erróneamente como insuficiente frente al 100 % teórico de una SLR manual, pero esta comparación ignora el costo operativo, según lo reportado en [13] donde se descargaron 10 556 artículos brutos, de los cuales 9 780 quedaron tras deduplicación. La lectura manual, a 1 minuto por registro, consumió un estimado de 163 horas netas para seleccionar 344 candidatos, de los cuales incluyeron 64.

El pipeline operó sobre 4 712 artículos, descartó el 82,9 % del corpus y produjo un pool de 807 candidatos en menos de 3 minutos, capturando 35 de los 64 GS (Recall 54,7 %, esfuerzo estimado de 13,4 h). La Tabla VI muestra que revisar solo los primeros 200 artículos (3,3 h) recupera el 29,7 % del GS, frente a las 163 h tradicionales.

Paradoja del Recall. Fenómeno mediante el cual un valor nominalmente bajo de Recall puede corresponder a un rendimiento superior al ponderarse por el esfuerzo cognitivo. Aquí, el 54,7 % se alcanza revisando solo el 17,1 % del corpus (807 documentos), mientras la SLR humana requirió el 100 %. La productividad relativa es entre 6,7 y 14,7 veces superior (35 GS/13,4 h y 19 GS/3,3 h frente a 64 GS/163 h). Un sistema que recupera "solo" la mitad de los relevantes puede ser más eficiente, porque el valor marginal del 45,3 % restante es desproporcionadamente costoso y, en muchos casos, su ausencia no compromete las conclusiones principales de la revisión.

Exhaustividad y techo teórico. Un Recall del 100 % requeriría un K que incluya todo el corpus, anulando la función de filtrado de las fases 1 y 2. Cada punto adicional de Recall exige un incremento desproporcionado del pool, con mayor tiempo de inferencia y dilución de la Precisión. La curva asintótica Recall@K (Fig. 2) confirma que, tras los 200 artículos superiores, las ganancias se vuelven marginales.

En consecuencia, el pipeline no sustituye el screening humano para máximo Recall, sino que acelera la Identificación

Primaria PRISMA concentrando la literatura de mayor relevancia en la cúspide del ranking. La reducción del esfuerzo se sitúa entre un estimado de 91,8 % (pool completo, 13,4 h) y 98,0 % (Top-200, 3,3 h), respecto a las 163 h reportadas en [13]. No busca reemplazar al revisor humano, sino actuar como filtro de alta densidad de relevancia.

C. Cumplimiento de PRISMA-trAIce

PRISMA-trAIce [14] es una guía de transparencia publicada en 2025 para documentar el uso de inteligencia artificial en revisiones sistemáticas. A diferencia de PRISMA-AI, enfocada en estudios sobre intervenciones de IA en salud, PRISMA-trAIce se aplica al uso de IA como método de síntesis de evidencia.

El pipeline cumple los ítems clave de PRISMA-trAIce [14]: (M2) identificación de herramientas (BM25, SBERT all-mpnet-base-v2, Cross-Encoder ms-marco-MiniLM-L-6-v2, open-source vía HuggingFace); (M3) aplicación exclusiva en Screening primario; (M4/M5) input título+abstract, output matriz con logits transformados a porcentajes de Relevancia Absoluta; (M6/M7) parámetros k_1=1,5, b=0,75, K=500 y thesis target; (M8) pipeline zero-shot autónomo con auditoría humana posterior; (M9/R2) métricas Recall@K, Precision@K, MRR, NDCG@K; (M10) ejecución local en Google Colab sin APIs terceros; (D1) limitaciones documentadas en Sección V, subsección D El ítem (R1) se cumple mediante la Fig. 4.

Para dar cumplimiento formal al ítem P-trAIce R1, el gráfico de la Fig. 4 mapea el flujo de filtrado sobre el layout visual estandarizado exigido por la extensión, distinguiendo explícitamente las exclusiones por reglas administrativas (fase ETL previa al screening) de las decisiones neuronales (IA).

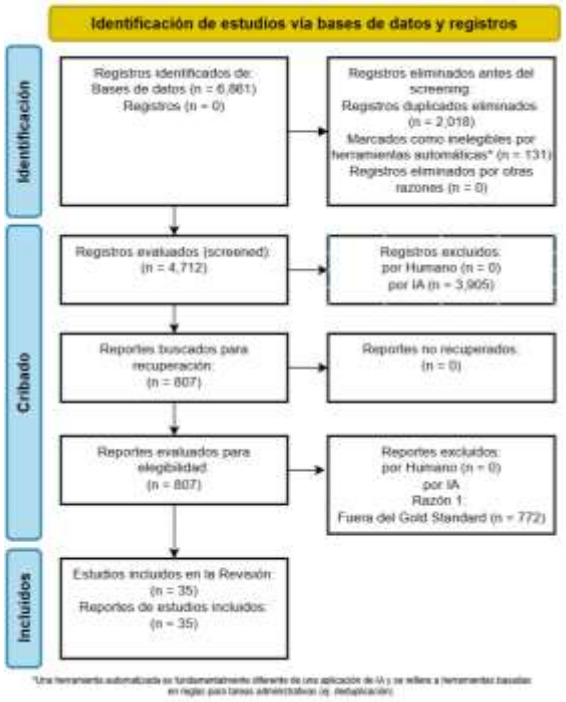


Fig. 4 Diagrama De Flujo Prisma-TrAIce Oficial (Ítem R1)

D. Limitaciones del estudio

Se identifican cinco limitaciones que delimitan la generalización de los resultados. Dominio único de validación: el Gold Standard procede de un solo dominio (VRPTW en investigación operativa). La transferibilidad del pipeline a dominios con menor terminología técnica estandarizada no ha sido verificada empíricamente. No obstante, la arquitectura fue concebida desde su diseño con una Taxonomía de 4 Motores: el modelo SBERT (all-mpnet-base-v2) y el Cross-Encoder (ms-marco-MiniLM) están optimizados para Ingeniería; dominios como Medicina requerirían modelos especializados (ej. BGE-M3 o BioLinkBERT), y dominios de Ciencias Sociales podrían beneficiarse de modelos multilingües.

Dependencia del thesis target: la calidad del ranking depende directamente de la ingeniería de prompts aplicada al thesis target. Un párrafo mal formulado degradará el rendimiento del Cross-Encoder sin que el sistema emita una alerta diagnóstica.

Umbral K empírico: el valor K = 500 se seleccionó empíricamente tras verificar que K = 100 produce resultados severamente degradados (Recall por debajo del 25 %). Un análisis formal de sensibilidad con K ∈ {100, 200, 300, 400, 500, 750, 1000} proporcionaría una curva de trade-off Recall-Precisión completa.

Dependencia del idioma del abstract: el pipeline opera exclusivamente sobre abstracts en inglés. Artículos publicados sin abstract, o con abstract en otros idiomas (español, portugués, chino), quedan automáticamente excluidos del corpus. Esta limitación es inherente al modelo lingüístico de SBERT y del Cross-Encoder, ambos pre-entrenados predominantemente en corpora anglohablantes.

Validación cruzada pendiente: para fortalecer la generalización, se requiere replicar el experimento sobre datasets de benchmarking multi-dominio como SYNERGY o los conjuntos estándar de ASReview. Esta replicación constituye el paso siguiente en la hoja de ruta de investigación, no una debilidad del presente estudio, que se posiciona como prueba de concepto validada en un dominio específico.

VI. Conclusiones

Se logró estructurar un corpus validado de 4 712 artículos replicando la estrategia del Gold Standard. El sistema descartó autónomamente el 82,9 % del ruido documental, reduciendo el esfuerzo frente al screening manual.

El pipeline opera con costo monetario cero sobre Google Colab (GPU T4) utilizando modelos open-source (licencia MIT), lo que demuestra que la automatización del screening no requiere inversión en infraestructura ni APIs comerciales. Esto resulta especialmente relevante para la comunidad académica latinoamericana, donde las restricciones presupuestarias constituyen una barrera estructural.

El sistema recuperó 35 de los 64 artículos del Gold Standard (Recall ajustado: 54,7 %), con un MRR de 1, en 2,4 minutos sobre GPU T4 frente a las 163 horas estimadas del screening manual reportado en [13]. La ablación confirmó la

irreductibilidad arquitectónica: eliminar BM25 redujo el Recall un 37,9 %, eliminar SBERT un 13,5 %, y eliminar el Cross-Encoder colapsó el NDCG@10 un 59,8 %.

Como trabajo futuro, se plantea la validación multi-dominio sobre datasets SYNERGY y ASReview, la integración de modelos long-context (nomic-embed-text-v1.5) para texto completo, y la incorporación de un ciclo opcional de active learning complementario.

AGRADECIMIENTO

Agradezco a la Universidad Católica Santo Toribio de Mogrovejo (USAT) y a la Facultad de Ingeniería por el respaldo académico y formativo brindado durante el desarrollo de esta investigación. El código fuente del pipeline se encuentra disponible públicamente bajo licencia MIT en: https://github.com/redentor159/pipeline-hibrido-screening-prisma

DECLARACIÓN DE USO DE HERRAMIENTAS DE IA

Se declara el uso de herramientas de IA generativa (Gemini 3.1 Pro, Claude Opus 4.6) como asistentes técnicos para la generación del código boilerplate en Python y la revisión gramatical, bajo el entorno de desarrollo Google Antigravity. La conceptualización, diseño experimental, análisis de resultados y conclusiones son obra original e íntegra de los autores humanos, quienes asumen la total responsabilidad intelectual del contenido publicado.

REFERENCIAS

[1] UNESCO, «Science Report: the race against time for smarter development - UNESCO Biblioteca Digital». Accedido: 29 de abril de 2026. [En línea]. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000377433

[2] M. J. Page *et al.*, «The PRISMA 2020 statement: an updated guideline for reporting systematic reviews», *BMJ*, vol. 372, p. n71, mar. 2021, doi: 10.1136/bmj.n71.

[3] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, y S. Ananiadou, «Using text mining for study identification in systematic reviews: a systematic review of current approaches», *Syst. Rev.*, vol. 4, n.º 1, p. 5, ene. 2015, doi: 10.1186/2046-4053-4-5.

[4] R. Borah, A. W. Brown, P. L. Capers, y K. A. Kaiser, «Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry», *BMJ Open*, vol. 7, n.º 2, p. e012545, feb. 2017, doi: 10.1136/bmjopen-2016-012545.

[5] Cochrane, «Manual Cochrane para revisiones sistemáticas de intervenciones (versión actual) | Cochrane». Accedido: 29 de abril de 2026. [En línea]. Disponible en: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current#gsc.tab=0

[6] G. V. Proaño *et al.*, «Effectiveness, barriers, and facilitators of overweight and obesity prevention strategies in Latin America; a scoping review and qualitative study in Colombia», *Lancet Reg. Health – Am.*, vol. 29, n.º p. 100656, ene. 2024, doi: 10.1016/j.lana.2023.100656.

[7] G. W. Furnas, T. K. Landauer, L. M. Gomez, y S. T. Dumais, «The vocabulary problem in human-system communication», *Commun ACM*, vol. 30, n.º 11, pp. 964-971, nov. 1987, doi: 10.1145/32206.32212.

[8] C. Lefebvre *et al.*, «Searching for and selecting studies», en *Cochrane Handbook for Systematic Reviews of Interventions*, Hoboken, NJ, USA: John Wiley & Sons, Ltd, 2019, pp. 67-107. doi: 10.1002/9781119536604.ch4.

[9] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», en *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, y T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019, pp. 4171-4186. doi: 10.18653/v1/N19-1423.

[10] R. van de Schoot *et al.*, «An open source machine learning framework for efficient and transparent systematic reviews», *Nat. Mach. Intell.*, vol. 3, n.º 2, pp. 125-133, feb. 2021, doi: 10.1038/s42256-020-00287-7.

[11] M. Ouzzani, H. Hammady, Z. Fedorowicz, y A. Elmagarmid, «Rayyan—a web and mobile app for systematic reviews», *Syst. Rev.*, vol. 5, n.º 1, p. 210, dic. 2016, doi: 10.1186/s13643-016-0384-4.

[12] J. Torres, C. Mulligan, J. Jorge, y C. Moreira, «PROMPTHEUS: A Human-Centered Pipeline to Streamline Systematic Literature Reviews with Large Language Models», *Information*, vol. 16, n.º 5, p. 420, may 2025, doi: 10.3390/info16050420.

[13] X. Liu, Y.-L. Chen, L. Y. Por, y C. S. Ku, «A Systematic Literature Review of Vehicle Routing Problems with Time Windows», *Sustainability*, vol. 15, n.º 15, p. 12004, ene. 2023, doi: 10.3390/su151512004.

[14] D. Holst, K. Moenck, J. Koch, O. Schmedemann, y T. Schüppstuhl, «Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist», *JMIR AI*, vol. 4, n.º 1, p. e80247, dic. 2025, doi: 10.2196/80247.

[15] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, y C. H. Schmid, «Semi-automated screening of biomedical citations for systematic reviews», *BMC Bioinformatics*, vol. 11, n.º 1, p. 55, ene. 2010, doi: 10.1186/1471-2105-11-55.

[16] K. E. K. Chai, R. L. J. Lines, D. F. Gucciardi, y L. Ng, «Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews», *Syst. Rev.*, vol. 10, n.º 1, p. 93, abr. 2021, doi: 10.1186/s13643-021-01635-3.

[17] C. F. Moreno-Garcia, C. Jayne, E. Elyan, y M. Aceves-Martins, «A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews», *Decis. Anal. J.*, vol. 6, p. 100162, mar. 2023, doi: 10.1016/j.dajour.2023.100162.

[18] J. Lin, R. Nogueira, y A. Yates, *Pretrained Transformers for Text Ranking: BERT and Beyond*. en Synthesis Lectures on Human Language Technologies. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-02181-7.

[19] J. A. Thankachan y B. Srinivasan, «An NLP Based Approach to Automate and Enhance the Systematic Review Within PRISMA Format», en *IFIP Advances in Information and Communication Technology*, A. Chandrabose, X. Fernando, y E. Mercier-Laurent, Eds., en Computer, Communication, and Signal Processing. Smart Solutions Towards SDG, vol. AICT-723. Chennai, India: Springer Nature Switzerland, mar. 2024, pp. 199-212. doi: 10.1007/978-3-031-73617-9_16.

[20] X. Mao, S. Zhuang, B. Koopman, y G. Zuccon, «Dense Retrieval with Continuous Explicit Feedback for Systematic Review Screening Prioritisation», en *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, jul. 2024, pp. 2357-2362. doi: 10.1145/3626772.3657921.

[21] C. Sharma, «Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers», 28 de mayo de 2025, arXiv:2506.00054. doi: 10.48550/arXiv.2506.00054.

[22] K. Song, X. Tan, T. Qin, J. Lu, y T.-Y. Liu, «MPNet: Masked and Permuted Pre-training for Language Understanding», en *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 16857-16867. Accedido: 2 de mayo de 2026. [En línea]. Disponible en: https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html

[23] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, y M. Zhou, «MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers», 6 de abril de 2020, arXiv:2002.10957. doi: 10.48550/arXiv.2002.10957.

[24] N. Reimers y I. Gurevych, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks», en *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, y X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, nov. 2019, pp. 3982-3992. doi: 10.18653/v1/D19-1410.

[25] R. Nogueira y K. Cho, «Passage Re-ranking with BERT», 14 de abril de 2020, arXiv:1901.04085. doi: 10.48550/arXiv.1901.04085.

[26] R. Ofori-Boateng, M. Aceves-Martins, N. Wiratunga, y C. Moreno-Garcia, «A Zero-Shot Monolingual Dual Stage Information Retrieval System for Spanish Biomedical Systematic Literature Reviews», en *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, y S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, jun. 2024, pp. 3725-3736. doi: 10.18653/v1/2024.naacl-long.206.

[27] S. Matysik, J. Wiśniewska, y P. K. Frankowski, «EmbedSLR: an open-source python framework for efficient embedding-based screening and bibliometric validation in systematic literature review», *SoftwareX*, vol. 32, p. 102416, dic. 2025, doi: 10.1016/j.softx.2025.102416.

[28] S. Robertson y H. Zaragoza, «The Probabilistic Relevance Framework: BM25 and Beyond», *Found. Trends Inf. Retr.*, vol. 4, n.º 1-2, pp. 1-174, sep. 2009, doi: 10.1561/1500000019.

[29] Elicit, «Elicit: AI for Scientific Research». Accedido: 8 de mayo de 2026. [En línea]. Disponible en: https://elicit.com/

[30] Consensus, «Consensus: AI for Research». Accedido: 8 de mayo de 2026. [En línea]. Disponible en: https://consensus.app/search/

[31] SciSpace, «SciSpace AI Research Agent | 150+ Tools, 280 M Papers». Accedido: 8 de mayo de 2026. [En línea]. Disponible en: https://scispace.com

[32] Covidence, «Pricing | Covidence - Better systematic review management», Covidence. Accedido: 29 de abril de 2026. [En línea]. Disponible en: https://www.covidence.org/pricing/

[33] DistillerSR, «DistillerSR: AI-Enabled Literature Review Software». Accedido: 8 de mayo de 2026. [En línea]. Disponible en: https://www.distillersr.com/