Desarrollo de un agente inteligente para recomendaciones médicas en embarazos mediante minería de texto y guías clínicas

Andrés Leonardo Mogollón Benavides

Asesor

Adriana del Pilar Noguera Torres

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Ingeniería Electrónica

2026

Resumen

La limitada disponibilidad de corpus de Preguntas y Respuestas (QA) validado en español para salud materna son un obstáculo para el desarrollo de sistemas de apoyo a decisiones clínicas. Este trabajo, dentro del Macroproyecto Minciencias 82244, construyó un corpus QA derivado de literatura biomédica de PubMed para entrenar modelos de Procesamiento de Lenguaje Natural (PLN) en entornos de telemedicina. Se implementó un *pipeline* que combinó extracción de artículos XML-JATS, segmentación semántica y vectorización con MiniLM-L12-v2. Los fragmentos obtenidos se indexaron en una base de datos vectorial (Chroma DB), garantizando la trazabilidad entre cada respuesta y su evidencia primaria. A partir de este proceso, se generaron 100 pares QA iniciales, los cuales se evaluaron mediante similitud coseno, obteniendo una media de 0.794 (IC 95%: [0.778, 0.811]) y 84% de pares clasificados como "Excelente" o "Bueno", constituyendo un recurso reproducible para sistemas RAG aplicados a seguimiento materno con recomendaciones basadas en evidencia.

***Palabras clave*:** Minería de Texto, Modelos Lingüísticos, Salud Aterna, Telemedicina, Información Biomédica.

Abstract

The limited availability of validated Spanish-language Question-Answer (QA) corpora for maternal health represents an obstacle to the development of clinical decision-support systems. This work, conducted within the Minciencias Macroproject 82244, constructed a QA corpus derived from biomedical literature in PubMed to train Natural Language Processing (NLP) models in telemedicine environments. A pipeline was implemented that combined XML-JATS article extraction, semantic segmentation, and vectorization using MiniLM-L12-v2. The resulting text fragments were indexed in a vector database (ChromaDB), ensuring traceability between each answer and its primary evidence source. One hundred QA pairs were generated and evaluated using cosine similarity, obtaining a mean of 0.794 (95% CI: [0.778, 0.811]) with 84% of pairs classified as "Excellent" or "Good," constituting a reproducible resource for Retrieval-Augmented Generation (RAG) systems applied to maternal monitoring with evidence-based recommendations.

***Keywords*:** Text Mining, Maternal Health, Telemedicine, Language Models, Biomedical Information

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Introducción

El avance de Grandes Modelos de Lenguaje o LLMs y su adaptación en telemedicina ofrecen nuevas oportunidades para la salud pública. Sin embargo, persisten desafíos en la atención oportuna y basada en evidencia en salud materna. Para abordarlos se requieren soluciones tecnológicas que garanticen precisión y trazabilidad de conocimiento clínico, no solo al acceso a la información.

El procesamiento del lenguaje natural ha estructurado información médica en contextos anglosajones, pero falta un corpus de preguntas y respuestas validado en español para salud materna que permita entrenar modelos de PLN. Este trabajo cubre la brecha mediante un corpus estructurado y reproducible derivado de literatura biomédica en PubMed, enfocándose específicamente en salud materna a diferencia de trabajos previos en español que adoptan enfoques generales.

Este estudio contribuye al Macroproyecto denominado *Agente inteligente basado en PLN para seguimiento materno en telemedicina pospandemia* avalado por el Ministerio de Ciencias Tecnología e Investigación Minciencias, Código 82244. Se plantea que la calidad de un corpus QA construido mediante minería de texto, segmentación semántica e indexación vectorial determina directamente la precisión del agente inteligente para telemedicina

En Colombia, la tasa de mortalidad materna fue 38.6 casos por cada 100.000 nacidos vivos en 2024 (Instituto Nacional de Salud, 2024), con disparidades regionales: en Chocó, Vichada y La Guajira superan 60 casos por 100.000. Aunque existen guías de práctica clínica y literatura biomédica disponible, su dispersión en múltiples repositorios limita su uso práctico en contextos clínicos con alta presión asistencial.

Trabajos previos han abordado parcialmente esta problemática: MOTHER proporciona corpus especializado en salud materna pero exclusivamente en inglés (Eyobu et al., 2025), mientras CasiMedicos-Arg ofrece validación en español pero en medicina general (Sviridova et al., 2024). Este trabajo propone integrar estas características: corpus QA en español especializado en salud materna. Se diferencia por tres aspectos: primero, utiliza artículos completos de PubMed como fuente (no solo resúmenes como BioASQ); segundo, implementa segmentación semántica sensible al contexto clínico (no fragmentación uniforme); tercero, valida computacionalmente mediante similitud coseno, proporcionando evaluación reproducible.

Este trabajo se estructura en cinco secciones principales: primero, la metodología describe el pipeline de extracción, segmentación y vectorización de 5.424 artículos de PubMed que resultaron en 284.031 fragmentos semánticamente indexados; segundo, el desarrollo documenta las nueve fases de procesamiento; tercero, los resultados reportan la evaluación de 100 pares QA mediante similitud coseno; cuarto, las conclusiones sinterizan contribuciones metodológicas, limitaciones y oportunidades de expansión del corpus.

Justificación

La construcción de corpus biomédicos de preguntas y respuestas en español representa una necesidad crítica en el ecosistema hispanohablante de inteligencia artificial aplicada a salud. Aunque existen corpus validados internacionalmente (BioASQ, MedExQA, MultiMedQA), estos se desarrollan exclusivamente en inglés, creando una brecha significativa para contextos clínicos hispanohablantes donde el acceso a información biomédica estructurada y validada sigue siendo limitado.

En Colombia, la mortalidad materna persiste como un problema de salud pública. Con tasas de 38.6 casos por 100.000 nacidos vivos a nivel nacional y concentraciones críticas en regiones como Chocó, Vichada y La Guajira (superiores a 60 casos por 100.000), existe una demanda inmediata de herramientas que cierren la brecha entre la evidencia clínica disponible en repositorios internacionales y su accesibilidad en contextos de telemedicina. Los sistemas de apoyo clínico basados en lenguaje natural representan una estrategia viable para este propósito.

Este proyecto justifica su implementación en tres dimensiones: primero, proporciona un corpus validado que habilita entrenamientos de modelos de procesamiento de lenguaje natural específicamente calibrados para el dominio materno-perinatal en español, cubriendo así una ausencia en la literatura. Segundo, demuestra metodologías reproducibles de minería de texto, segmentación semántica e indexación vectorial que pueden generalizarse a otros dominios críticos (pediatría, oncología, enfermedades infecciosas), multiplicando el impacto más allá de este caso de uso. Tercero, genera infraestructura operacional (corpus indexado en Chroma DB con 284031 embeddings) directamente integrable en arquitecturas RAG y sistemas de telemedicina, transitando desde investigación hacia operacionalización.

El macroproyecto Minciencias 82244 requiere insumos estructurados para entrenar agentes inteligentes en telemedicina pospandemia. Este trabajo proporciona tanto el corpus validado como la metodología para su construcción, eliminando un cuello de botella crítico en la cadena de desarrollo. Además, este esfuerzo fortalece la capacidad local en ingeniería de datos biomédicos, demostrando cómo la ingeniería electrónica contribuye directamente a soluciones de salud pública.

Finalmente, la evaluación computacional mediante similitud semántica establece un estándar de validación automatizada que reduce dependencia de juicio experto para iteraciones posteriores, acelerando ciclos de mejora del corpus y haciéndolo escalable. En síntesis, el presente trabajo resuelve una brecha tecnológica documentada, habilita un macroproyecto institucional y demuestra viabilidad metodológica para un ecosistema hispanohablante de IA en salud.

Planteamiento del Problema

Uno de los indicadores clave en el desarrollo de un país es la tasa de mortalidad materna, refleja el estado de los sistemas de salud, brechas sociales y brechas económicas de su población. En Colombia, el Instituto Nacional de Salud (2024), reportó 38.6 casos por cada 100000 nacidos vivos, una reducción frente a décadas anteriores. La situación se agrava en territorios como el Chocó, Vichada o La Guajira, donde las tasas superan los 60 casos por cada 100000, evidenciando profundas desigualdades regionales. La hipertensión, las hemorragias y eventos tromboembólicos continúan siendo responsables de dos tercios de estas muertes.

Aunque existen guías de práctica clínica y literatura biomédica que proporcionan evidencia sólida para la prevención y tratamiento de complicaciones maternas (como los protocolos del Ministerio de Salud 2013), su extensión, organización y dispersión en múltiples repositorios limitan su uso práctico en contextos clínicos con alta presión asistencial. Una revisión Cochrane (Agarwal et al., 2025) demostró que herramientas digitales mejoran las decisiones clínicas y la adherencia a protocolos, pero en una urgencia obstétrica, localizar una recomendación específica entre cientos de páginas es inviable. Esta carencia ha impulsado el desarrollo de corpus de preguntas y respuestas biomédicas para entrenar modelos de lenguaje que hagan accesible la evidencia. Sin embargo, la mayoría de estos recursos fueron creados en inglés y carecen de especialización en salud materna en español.

Este problema lo enfrentan directamente los profesionales de salud que atienden gestantes en hospitales y centros de atención, e indirectamente las mujeres embarazadas, al recibir diagnósticos tardíos o decisiones clínicas sin respaldo en evidencia.

La mayoría de corpus biomédicos existentes se desarrollan en el idioma inglés: BioASQ, MedExQA y MultiMedQA (Krithara et al., 2022; Singhal et al., 2023; Kim et al., 2024). Algunas

propuestas multilingües como CasiMedicos-Arg incorporan español, pero en escala reducida (Sviridova et al., 2024). En obstetricia, los recursos son aún más limitados: MOTHER contiene 503 pares QA en inglés, y Pregnant Questions carece de trazabilidad textual hacia evidencia científica (Eyobu et al., 2025; Srikanth et al., 2023). Esta situación evidencia la ausencia de un corpus QA biomédico en español especializado en salud materna, construido desde artículos completos y evaluado con métricas computacionales que garanticen coherencia semántica. El proyecto cierra esta brecha mediante minería de PubMed y generación de pares QA en español con modelos preentrenados, produciendo un recurso con trazabilidad hacia fuentes originales.

El problema no es la falta de literatura biomédica, sino la ausencia de un corpus estructurado en español que permita entrenar modelos especializados en salud materna. Este proyecto ofrece una respuesta ingenieril viable, demostrando cómo desde la ingeniería electrónica orientada hacia las líneas de industrias inteligentes aplicada a ciencia de datos puede contribuir a la innovación tecnológica en salud. El planteamiento se articula con la macrolínea de Desarrollo económico y social basado en ciencia y tecnología, las líneas de Ciencia de Datos y Transformación Digital Inteligente, y el ODS 9 sobre innovación e infraestructura.

Objetivos

Objetivo General

Construir un corpus biomédico de preguntas y respuestas en español especializado en salud materna, a partir de la minería de texto, que pueda servir como insumo de entrenamiento y evaluación para modelos de procesamiento de lenguaje natural en el dominio médico

Objetivos Específicos

Diseñar un pipeline de procesamiento y post - procesamiento de datos para la extracción de información biomédica y la generación de pares pregunta-respuesta en español.

Establecer una estructura de base de datos que permita el almacenamiento eficiente, organizado y trazable del corpus QA generado.

Implementar un proceso de evaluación computacional del corpus mediante métricas automáticas que aseguren la coherencia semántica de los pares pregunta-respuesta

Estado del Arte

Los corpus de pregunta-respuesta han mantenido su popularidad por la capacidad de servir como base del entrenamiento de modelos de lenguaje (De Ingeniería del Conocimiento, 2025). Estos recursos requieren, más allá de volumen y diversidad, metodologías que garanticen validez y aplicabilidad específica. Este capítulo analiza los avances y desafíos en cuatro áreas: metodologías de construcción de corpus QA, corpus biomédicos existentes, minería de texto en literatura y PLN en español para medicina.

Metodologías de Construcción de Corpus QA

La construcción de corpus QA involucra tres procesos centrales: generación de preguntas, validación clínica y definición de métricas de evaluación. En la generación de preguntas, la literatura contrasta dos enfoques: basados en plantillas y métodos neuronales. Los primeros, como EHRXQA, formula preguntas consistentes alineadas con datos estructurados (tablas clínicas e imágenes radiológicas) mediante esquemas diseñados por expertos (Bae et al., 2023). Si bien aseguran precisión, su alcance es limitado: las plantillas siguen patrones gramaticales fijos que no capturan la variabilidad de consultas reales, además su creación requiere tiempo y esfuerzo del personal médico. Los métodos neuronales, como ACS-QG basado en GPT-2 y modelos T5, ofrecen mayor flexibilidad y naturalidad en la formulación desde texto libre (Liu, Wei, Niu, Chen, & He, 2020). Sin embargo, introducen riesgo de contenido clínicamente incorrecto: los LLM pueden generar afirmaciones coherentes pero no fácticas o perjudiciales, fenómeno conocido como alucinación, limitación crítica en contextos médicos.

El uso de ontologías biomédicas como UMLS o MeSH ha buscado mitigar estos problemas mediante normalización terminológica y coherencia semántica (Krithara et al., 2023). No obstante, su diseño centrado en inglés y contextos anglosajones dificulta la adaptación a

escenarios multilingües, donde variaciones terminológicas regionales complican la estandarización de términos clínicos.

En validación, no existe consenso único. Diferentes recursos emplean métodos variados: BioASQ recurre a expertos biomédicos (Krithara et al., 2022), MedMCQA se fundamenta en exámenes médicos, mientras MedExQA y MultiMedQA utilizan evaluaciones humanas. Sin embargo, estudios como el de Srikanth et al. (2024) señalan dificultades para que especialistas logren acuerdo al clasificar información. En evaluación, métricas automáticas como BLEU y ROUGE siguen siendo ampliamente utilizadas por ser prácticas ante ausencia de validación clínica a escala. Proyectos como MedExQA han complementado estas con medidas semánticas sofisticadas como BERTScore, mientras MultiMedQA enfatiza la necesidad de validación experta para garantizar seguridad clínica (Singhal et al., 2023).

Corpus Biomédicos Existentes

Una vez revisadas estas metodologías, es pertinente analizar los corpus biomédicos existentes. BioASQ ha establecido un estándar de trazabilidad al vincular preguntas con fragmentos específicos de resúmenes de PubMed, conectando cada respuesta directamente con su fuente mediante marcadores que identifican la ubicación exacta del texto (Krithara et al., 2022). MedExQA, en contraste, sacrifica escala para priorizar explicaciones múltiples por cada par pregunta-respuesta, enfatizando calidad del razonamiento sobre cobertura masiva (Kim et al., 2024).

EHRXQA integra registros clínicos estructurados con imágenes radiológicas, habilitando razonamiento multimodal, pero sin ofrecer trazabilidad textual mediante spans (Bae et al., 2023). CasiMedicos-Arg aporta un enfoque multilingüe y argumentativo al incluir

explicaciones médicas anotadas en español, inglés, francés e italiano, permitiendo evaluar no solo respuestas finales sino la lógica subyacente en justificaciones (Sviridova et al., 2024).

En salud materna, los avances son más limitados. MOTHER contiene 503 pares QA derivados de encuestas en Uganda, constituyendo un esfuerzo pionero validado por profesionales, aunque limitado en escala, idioma y trazabilidad hacia literatura científica (Eyobu et al., 2025). Pregnant Questions explora la dimensión pragmática de consultas maternas, mostrando que muchas preguntas incluyen supuestos implícitos que requieren interpretaciones profundas (Nygaard et al., 2023).

MultiMedQA integra múltiples corpus en un mismo benchmark, permitiendo evaluar modelos en diversas tareas, aunque hereda limitaciones de sus datasets base y carece de consistencia en trazabilidad de evidencia (Singhal et al., 2023). En conjunto, estos corpus reflejan un campo en expansión pero fragmentado en enfoques y limitado en dominios especializados como el materno.

Minería de Texto en la Literatira Biomédica

PubMed Central usa el formato JATS XML para almacenar artículos biomédicos completos. Este formato permite extraer metadatos, resúmenes y secciones mediante herramientas como JATSdecoder (Boschen, 2021). La minería de texto identifica entidades biomédicas (genes, enfermedades, medicamentos) usando librerías como scispaCy y modelos especializados como BioBERT (Neumann et al., 2019; Lee et al., 2020). Un componente clave es la segmentación o chunking de textos científicos, necesaria para dividir documentos extensos en unidades coherentes que faciliten la generación de pares QA. La segmentación tradicional enfrenta dificultades en el lenguaje biomédico, debido a abreviaturas, compuestos y estilos de citación particulares. Por ello, se ha avanzado hacia chunking semántico, donde representaciones

contextuales basadas en embeddings (como SciBERT o BioBERT) permiten crear fragmentos más significativos para sistemas de QA. En este sentido, herramientas como scispaCy han demostrado segmentación robusta y adaptada al dominio, lo que resulta crucial para proyectos que, como este, se basan en artículos completos de PubMed y requieren asegurar coherencia clínica en cada segmento antes de generar preguntas y respuestas.

PNL en Español para Medicina

La mayoría de corpus biomédicos han sido desarrollados exclusivamente en inglés, lo que limita su aplicabilidad en contextos globales. CasiMedicos-Arg constituye una excepción al ofrecer datos multilingües con explicaciones médicas anotadas, incluyendo el español, aunque su escala sigue siendo reducida (Sviridova et al., 2024). MOTHER aporta un corpus específico de salud materna, pero únicamente en inglés (Eyobu et al., 2025). Modelos multilingües como mBERT, XLM-RoBERTa y Medical mT5 han mostrado capacidad para transferir conocimientos entre idiomas, y estrategias como la proyección de etiquetas seguida de revisión manual han probado ser efectivas para construir datasets multilingües. Sin embargo, la investigación coincide en que la falta de benchmarks robustos en español constituye un obstáculo para avanzar hacia sistemas de QA biomédicos confiables en este idioma. La salud materna, además, presenta necesidades lingüísticas específicas: términos como "presión alta" frente a "hipertensión gestacional" ilustran cómo la diversidad terminológica puede impactar directamente en la interpretación clínica.

Ante este panorama de avances y limitaciones, emerge un patrón claro: existen avances notables como sofisticación en metodologías de generación de preguntas, explicaciones argumentativas y multimodales, y desarrollo de recursos multilingües incipientes. Sin embargo, persisten limitaciones críticas: dependencia del inglés, falta de validación clínica escalable,

escasa integración de artículos completos de PubMed y ausencia de corpus QA en salud materna con trazabilidad verificable. Estos vacíos justifican el presente proyecto, que busca construir un corpus QA biomédico a partir de artículos completos de PubMed para entrenar modelos, con énfasis en español y adaptación a salud materna.

Metodología

Se implementó un enfoque cuantitativo-experimental bajo el modelo CDIO, priorizando decisiones técnicas evaluadas empíricamente. La selección de herramientas respondió a criterios específicos: (1) soporte robusto para dominio biomédico (APIs de NCBI, librerías especializadas), (2) capacidad multilingüe dada la brecha en corpus QA español-medicina, (3) procesamiento local mediante computación de alto rendimiento disponible, evitando dependencias de APIs de costo elevado, (4) validación automatizada ante ausencia de validadores clínicos expertos en la fase de implementación

Minería de Texto

Se utilizó la API Entrez del National Center for Biotechnology Information NCBI, implementada mediante la librería BioPython, para automatizar la búsqueda y extracción de literatura biomédica especializada en salud materna desde PubMed. La ecuación de búsqueda abarcó conceptos generales en salud materna y complicaciones obstétricas principales, incluyendo publicaciones de acceso abierto entre 2022 y 2025.

Se extrajeron artículos con identificador PMCID válido e implementó un proceso de descarga automatizada de archivos XML JATS desde los servidores de Europe PMC, con acceso secundario a NCBI PMC. Las descargas se ejecutaron en paralelo mediante programación multihilo para optimizar tiempos.

Los archivos XML se validaron mediante BeautifulSoup para confirmar la presencia de texto completo verificando la etiqueta <body>. De cada registro se extrajeron título, resumen y párrafos, descartando referencias, agradecimientos, información de autoría y material suplementario. El contenido se organizó en dos archivos CSV: uno de resumen (pmc_jats_resumen.csv) diferenciando artículos con cuerpo completo de aquellos con solo

metadatos, y otro de contenido procesado (pmc_jats_contenido_limpio.csv) con cada fragmento identificado por sección, tipo de elemento y posición.

Generación de Chunks y Embeddings

Se segmentó el corpus aplicando chunking semántico, necesario para estructurar información en unidades lógicamente coherentes sin cortes abruptos que fragmenten el sentido clínico (Nayak, 2024). Un chunking basado únicamente en tamaño fijo habría generado fragmentos desconectados en contextos biomédicos con terminología especializada. Se empleó SemanticChunker de LangChain, que divide textos mediante variaciones semánticas detectadas por embeddings, capturando límites naturales en el contenido médico.

Se seleccionó el modelo MiniLM (Paraphrase-multilingue-MiniLM-L12_v2) por su equilibrio entre eficiencia computacional y capacidad multilingüe, permitiendo procesamientos locales en la infraestructura disponible sin comprometer la captura semántica. Los documentos se procesaron sección por sección, generando un archivo intermedio donde cada fila correspondía a un fragmento identificado con artículo de origen, sección y longitud.

Posteriormente se aplicó limpieza y re-chunking para controlar extensión: fragmentos menores de 100 palabras fueron descartados por insuficiencia representativa de conceptos clínicos, y los excesivamente largos se dividieron en subunidades, manteniendo distribución homogénea de longitudes para evitar sesgos en la generación posterior de pares QA.

La vectorización semántica se implementó mediante MiniLM en SentenceTransformers, transformando cada chunk en un vector que preserva su información semántica, generando un corpus segmentado, normalizado y representado en espacio vectorial.

Indexación a Bases de Datos Vectorial

Se empleó Chroma DB, un almacén de vectores de código abierto que permite almacenar

y recuperar embeddings junto con metadatos para su posterior uso en modelos de lenguaje de gran tamaño. (datacamp, 2024).

Los vectores generados en la fase previa (archivo pmc_chunks_embeddings_sbert.csv) se cargaron y organizaron en colecciones, asignando a cada chunk un identificador único que combina el documento de origen y su posición en el corpus.

Los registros se indexaron en lotes de mil entradas para equilibrar eficiencia computacional con manejo robusto de la gran cantidad de instancias generadas. A cada vector se asociaron metadatos estructurados (sección del documento, identificador del artículo, longitud del fragmento) esenciales para vincular respuestas generadas posteriormente con su evidencia de origen, requisito crítico en contextos clínicos.

Se implementó persistencia de la colección para evitar repetir minería, limpieza, chunking y generación de embeddings en iteraciones posteriores, permitiendo reutilización eficiente del corpus vectorizado y reproducibilidad de la etapa de generación de pares QA.

Generación de Pares Pregunta-Respuesta

Consideraciones Lingüísticas

Si bien los fragmentos biomédicos del corpus se encuentran en inglés (idioma original de las publicaciones en PubMed Central), la generación de pares pregunta-respuesta se realizó en español. Este enfoque asimétrico fue necesario para aprovechar la completitud de la literatura biomédica en inglés sin comprometer la utilidad del recurso para la comunidad hispanohablante. Traducir el corpus completo habría introducido imprecisiones terminológicas tempranas que se propagarían en todas las fases posteriores.

Se seleccionó el modelo gpt-oss:20b por su capacidad de procesamiento local sin dependencias de APIs comerciales, garantizando control sobre generación y costos. El modelo

fue instruido para formular preguntas en español basándose en comprensión del contenido en inglés, evitando traducción mecánica que reduciría naturalidad de las preguntas. Las respuestas se generaron mediante traducción fiel del fragmento fuente sin agregar información externa, asegurando que cada respuesta permanezca anclada en la evidencia original.

Procedimiento de Generación

Como ejecución inicial se generó una muestra de 100 pares QA. Aunque el corpus disponible contiene aproximadamente 2 GB de contenido procesado, esta primera ejecución sirve de validación antes de un escalado a volúmenes mayores. Se implementó un módulo de filtrado temático para priorizar fragmentos con relevancia clínica explícita en salud materna y perinatal, evitando ruido de contenido tangencialmente relacionado. La estrategia utilizó una lista ampliada de palabras clave bilingüe (parto, embarazo, puerperio, complicaciones obstétricas, salud neonatal) diseñada a partir de terminología de guías clínicas nacionales. El sistema aplicó normalización mediante conversión a minúsculas y singularización de palabras para capturar variaciones lingüísticas.

La generación se ejecutó mediante un pipeline de dos pasos usando gpt-oss:20b implementado localmente con Ollama. Una arquitectura de dos fases permitió separación clara entre generación de pregunta y respuesta, mejorando control de calidad. En el primer paso se generó una pregunta en español cuya respuesta debía fundamentarse en el contenido del texto origen, evitando inferencias no respaldadas. En el segundo paso se generó la respuesta mediante traducción fiel del fragmento al español sin información externa.

El proceso incorporó límite de 20 intentos máximos, timeout de 180 segundos por llamada, y guardado incremental cada 50 pares para garantizar persistencia. Los pares se

almacenaron en CSV con metadatos de trazabilidad hacia documento fuente, sección e identificador del chunk.

Validación y Métricas de Calidad Pares Pregunta-Respuesta

La evaluación de coherencia semántica se realizó mediante similitud coseno, métrica apropiada para contextos donde fragmentos fuente (inglés) y respuestas generadas (español) residen en idiomas distintos. Se empleó el modelo paraphrase-multilingual-MiniLM-L12-v2 de Sentence Transformers, capaz de mapear textos a un espacio vectorial denso de 384 dimensiones con soporte multilingüe (Keithhon, 2022), permitiendo capturar si la respuesta en español preserva fielmente el contenido del fragmento en inglés.

Se calculó similitud coseno para cada par, obteniendo valores en [0,1]. Los pares se clasificaron según umbrales definidos empíricamente mediante inspección manual de una muestra de 20 pares iniciales: correspondencia literal ($\geq$0.75, indicando alta fidelidad semántica y traducción cercana), paráfrasis fiel (0.55-0.74, preservación del contenido central pero con reformulación), discordante ($<$0.55, baja correspondencia sugiriendo generación parcialmente externa al fragmento). Este procedimiento permitió calibrar umbrales a la realidad específica del corpus biomédico-español.

Se calcularon medidas descriptivas: media, mediana, desviación estándar, rango, percentiles 25 y 75, e intervalos de confianza 95% para la media. Se cuantificaron frecuencias absolutas y relativas por categoría de clasificación.

Los resultados se visualizaron mediante histograma con líneas de referencia en umbrales y media, gráfico de barras con distribución por categorías, e identificación de cinco casos con mayor y menor similitud para análisis cualitativo de patrones de error.

Limitaciones Metodológicas

Este enfoque presenta limitaciones inherentes al uso de modelos de lenguaje para generación automática. La calidad de los pares depende de la capacidad del modelo para comprender terminología biomédica especializada, donde conceptos complejos pueden ser malinterpretados. Además, la validación automatizada mediante similitud coseno mide coherencia semántica pero no seguridad clínica del contenido.

Desarrollo

El desarrollo experimental se ejecutó en un entorno local con sistema operativo Linux y aceleración por GPU NVIDIA, optimizando tiempos de ejecución para generación de embeddings y modelos de lenguaje. Se empleó Python 3 como lenguaje principal por su extensibilidad y ecosistema de bibliotecas para ciencia de datos (Van Rossum & Drake, 2009).

Los módulos se implementaron en Jupyter Notebook, entorno interactivo que integra código, visualizaciones y documentación en flujos reproducibles (Kluyver et al., 2016). Las librerías se instalaron mediante pip, garantizando compatibilidad de dependencias y trazabilidad experimental.

La Tabla 1 resume las librerías y módulos empleados en el pipeline, cubriendo minería de texto biomédico, segmentación semántica, vectorización y evaluación de coherencia semántica.

Tabla 1

Librerías y Módulos Usados en el Entorno de Desarrollo

Librería / módulo	Función principal	Referencia
xml.etree.ElementTree	Parseo de archivos XML para extraer secciones de artículos científicos.	Van Rossum & Drake (2009)
Bio.Entrez (*BioPython*)	Conexión con la API Entrez de PubMed para búsqueda y descarga de artículos.	Cock et al. (2009)
BeautifulSoup4	Limpieza y análisis estructural de archivos XML JATS.	Richardson (2007)
Pandas	Manipulación de datos tabulares mediante *Dataframe*.	Reback et al. (2020)
Requests	Gestión de solicitudes HTTP para descarga de documentos.	Chandra (2018)
concurrent.futures	Ejecución paralela de tareas de descarga (*multithreading*).	Van Rossum & Drake (2009)
tqdm	Seguimiento de progreso en iteraciones largas.	da Costa-Luis (2019)
langchain_experimental text_spliter.Semantic Chunker	Segmentación semántica de texto para generación de *chunks* coherentes	
	Tokenización y representación contextual de texto.	Chase (2023)

Librería / módulo	Función principal	Referencia
transformers (AutoTokenizer, AutoModel)	Generación de *embeddings* multilingües y cálculo de similitud.	Wolf et al. (2020)
sentence_transformers		Reimers & Gurevych (2019)
chromadb	Almacenamiento vectorial e indexación semántica persistente.	Chroma (2023)
torch (*PyTorch*)	Computación tensorial acelerada en GPU para embeddings y LLMs.	Paszke et al. (2019)
numpy	Operaciones numéricas básicas y manipulación de matrices.	Harris et al. (2020)
unicodedata, re, json, os, datetime, random	Módulos estándar de Python para normalización, expresiones regulares y manejo de archivos.	Van Rossum & Drake (2009)

Nota. Las herramientas seleccionadas son de código abierto y permiten ejecutar cada etapa del pipeline de forma local: búsqueda bibliográfica (Bio.Entrez), procesamiento XML (BeautifulSoup), segmentación semántica (LangChain), vectorización (Sentence Transformers) e indexación (ChromaDB), sin dependencias de APIs comerciales. Adaptado de. *Autoría propia.*

Minería y Procesamiento de Contenido Biomédico

Búsqueda Automatizada

Se implementó un sistema automatizado de búsqueda mediante la API Entrez del National Center for Biotechnology Information NCBI, integrada a través de BioPython, para ejecutar consultas sistematizadas hacia PubMed y acceder a artículos sin intervención manual.

La ecuación de búsqueda combinó operadores booleanos AND/OR en tres dimensiones interconectadas:

Conceptos Generales. términos como "maternal health" OR "pregnancy" en campos de título y resumen (TIAB), asegurando coincidencia en metadatos de alta relevancia.

Complicaciones Específicas. términos especializados sobre morbilidad materna severa (preeclampsia, eclampsia, hemorragia postparto, sepsis, infección materna, obstrucción del parto, ruptura uterina, hipertensión gestacional, diabetes gestacional).

Restricciones: publicaciones de 2022 a 2025 en acceso abierto ("free full text").

Figura 1

Ecuación de Búsqueda Booleana en PubMed para Salud Materna

```
query = """(
("maternal health"[TIAB] OR "pregnancy"[TIAB])
AND (
    "preeclampsia"[TIAB] OR "eclampsia"[TIAB] OR
    "postpartum hemorrhage"[TIAB] OR "obstetric hemorrhage"[TIAB] OR
    "sepsis"[TIAB] OR "maternal infection"[TIAB] OR
    "obstructed labor"[TIAB] OR "uterine rupture"[TIAB] OR
    "hypertensive disorders of pregnancy"[TIAB] OR
    "gestational diabetes"[TIAB] OR
    "maternal morbidity"[TIAB] OR "maternal mortality"[TIAB]
)
AND ("2022/01/01"[PDAT] : "2025/12/31"[PDAT])
AND ("free full text"[FILT])
)"""
```

Nota. Se empleó la nomenclatura estándar de PubMed: TIAB (Title and Abstract), PDAT (Publication Date), FILT (Filters). Adaptado de. *Autoría propia.*

Ejecución de la Búsqueda

La consulta se ejecutó mediante la función Entrez.search() de BioPython con el parámetro usehistory="y", que permite almacenar resultados en servidores remotos. Este parámetro retorna dos tokens clave: WebEnv (identificador de sesión remota) y QueryKey (identificador para la ecuación), permitiendo recuperación posterior sin limitación de cantidad (Chang et al., 2009). NCBI guarda la búsqueda y devuelve una referencia a los resultados, permitiendo consultas anticipadas y cachéo temporal para mejorar velocidad.

Extracción de Identificadores PMCID Válidos

Los 13.290 artículos retornados estaban identificados por PMID (PubMedIDs), pero las fases posteriores requieren formato XML JATS desde PubMed Central, que usa identificadores diferentes: PMCID (PubMed Central IDs). No todos los artículos indexados en PubMed poseen

PMCID, por lo que fue necesario mapear los PMID hacia sus correspondientes PMCID para determinar cuántos artículos podían procesarse.

Se recuperaron los 13.290 artículos mediante Entrez.efetch() en lotes de 200 registros. Para cada lote, la función retorna XML con metadatos que incluyen múltiples identificadores. Se implementó un parser con xml.etree.ElementTree que navegó la estructura XML buscando elementos <ArticleId> con atributo IdType="pmc", extrayendo el PMCID único de cada artículo. Solo artículos con PMCID válido fueron registrados; los demás fueron descartados.

Los PMCID extraídos se almacenaron en un dataframe df_pmc y se eliminaron duplicados mediante drop_duplicates().

Tabla 2

Resultados de Extracción de Identificadores PMCS Válido

Métrica	Cantidad	Porcentaje
Artículos recuperados en lotes PMID	13290	100%
PMCID validados extraídos (sin duplicados)	7728	58.1%
Artículos sin PMCID en PubMed central	5562	41.9%

Nota. El 58.1% de artículos con PMCID refleja que aproximadamente 41.9% carecía de identificador en PubMed Central, situación esperada ya que PubMed es un repositorio bibliográfico más amplio que PubMed Central. Los 7.728 PMCID válidos fueron almacenados en pmc_index.csv para la siguiente etapa de descarga de archivos XML. Adaptado de. *Autoría propia.*

Recuperación Paralela de Archivos XML JATS

Se implementó un sistema de descarga multihilo para los 7.728 archivos XML JATS, utilizando concurrent.futures.ThreadPoolExecutor configurado con 5 hilos de trabajo para

optimizar tiempos sin bloquear la ejecución.

El sistema empleó una estrategia dual de fuentes con fallback automático: intenta primero descargar desde Europe PMC (EBI), y si falla con error 404 o conexión, redirige a NCBI PMC (NCBI). Solo si ambas fuentes son inaccesibles se registra el PMCID como error.

Configuración de Validación y Tolerancia a Fallos. Timeout: 30 segundos máximo por solicitud HTTP. Validación de contenido: búsqueda de etiqueta <article> en el XML descargado para confirmar estructura JATS válida. Verificación de existencia: antes de descargar, verifica si el archivo ya existe localmente (estado "skipped"). Almacenamiento: archivos guardados en directorio pmc_xml/.

Ejecución Paralela. Los 7.728 PMCID se distribuyeron entre 5 workers que procesaron descargas de forma independiente. El sistema utilizó as_completed() para procesar cada descarga apenas terminaba, sin esperar a los otros, permitiendo contabilizar resultados en tiempo real, liberar memoria inmediatamente y optimizar uso de recursos.

Tabla 3

Resultados de Descarga Paralela de Archivos XML JATS desde PubMed Central PMC

Métrica	Cantidad	Porcentaje
PMCID válidos disponibles para descarga	7728	100%
Archivos descargados exitosamente	7635	98.8%
Archivos omitidos (ya existían)	0	0%
Descargas fallidas (ambas fuentes inaccesibles)	93	1.2%
Total procesado	7728	100%

Nota. De los 7.728 PMCID válidos, se descargaron 7.635 archivos exitosamente (98.8%). Europe PMC proporcionó 6.142 archivos (80.4%) como fuente principal, mientras que NCBI PMC aportó 1.493 (19.6%) como fuente de respaldo. Noventa y tres descargas fallaron (1.2%) cuando ambas fuentes resultaron inaccesibles o los XML no contenían el elemento <article>. Los

7.635 archivos válidos fueron almacenados en el directorio pmc_xml/ para procesamientos posteriores. Adaptado de. *Autoría propia.*

Clasificación de Contenido por Elementos JATS

PubMed Central almacena tanto artículos de acceso abierto completo como registros con solo metadatos (título, autores, resumen). Para diferenciar ambas categorías, se implementó un proceso que identificaba la presencia del elemento <body> como criterio determinante: su presencia indica texto completo, su ausencia indica solo metadatos bibliográficos.

Se utilizó BeautifulSoup para parsear los archivos XML buscando:

- PMCID
- <body>
- <article-title>
- <abstract>

Se implementó manejo de excepciones para archivos que no parseaban correctamente, registrándose como has_body=False (metadatos únicamente).

Tabla 4

Distribución de Artículos por Presencia de Elemento

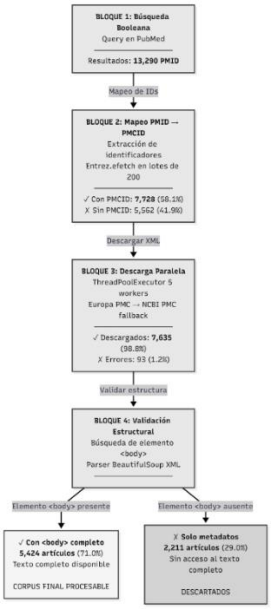
Categoría	Cantidad	Porcentaje	Descripción
Artículos con elemento <body> (texto completo)	5424	71.0%	Acceso abierto completo; contienen el cuerpo del artículo.
Artículos sin elemento <body> (metadatos únicamente)	2211	29.0%	Solo metadatos bibliográficos disponibles.
Total archivos validados	7635	100%	----

Nota. El 71.0% de artículos incluye <body> completo. El 29.0% restante contiene solo metadatos debido a restricciones editoriales, embargos o políticas sin acceso abierto, proporción común en PubMed Central que refleja la diversidad de licencias entre revistas. Los resultados se registraron en pmc_jats_resumen.csv, documentando presencia de <body>, título y primeros 300

caracteres del resumen. Los 5.424 artículos con texto completo constituyeron la entrada para las siguientes etapas del pipeline. Adaptado de. *Autoría propia.*

Figura 2

Pipeline de Minería de Texto, Identificación, Descarga y Validación de Artículos



Nota. El diagrama ilustra los cuatro primeros bloques del pipeline de minería de texto. El Bloque 1 retorna 13.290 artículos de PubMed identificados por PMID. El Bloque 2 mapea estos PMID a PMCID, resultando en 7.728 identificadores válidos (58.1%) para acceder a PubMed Central. El Bloque 3 ejecuta descargas paralelas mediante 5 workers con estrategia de fallback (Europa PMC primario, NCBI PMC secundario), obteniendo 7.635 archivos XML (98.8%). El Bloque 4 valida la presencia del elemento <body>, identificando 5.424 artículos con texto completo (71.0%) y 2.211 con solo metadatos (29.0%). Adaptado de. *Autoría propia.*

Parseo y Limpieza

Se implementó un procedimiento de parseo y limpieza para transformar únicamente los 5.424 artículos con texto completo en una estructura tabular. El propósito fue descomponer cada

artículo en unidades procesables (párrafos y tablas), eliminar ruido editorial y preservar trazabilidad hacia el documento de origen.

Se procesaron artículos con <body> completo, segmentando su contenido en elementos lógicos y descartando secciones no clínicas como referencias, agradecimientos, notas de autoría y resúmenes.

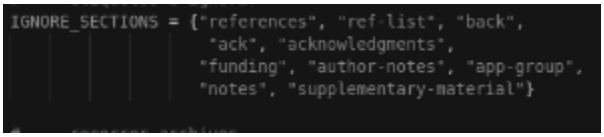
Filtrado de Ruido Editorial

Uno de los desafíos de este proyecto fue eliminar contenido no clínico de la estructura XML. Se estableció un conjunto de secciones a ignorar, identificadas mediante el atributo sec-type: references, ref-list, back, ack, acknowledgments, funding, author-notes, app-group, notes, supplementary-material.

El procedimiento inició con la lectura de pmc_jats_resumen.csv, que contenía los PMCID con <body> completo. Para cada identificador se localizó su XML correspondiente y se parseó con BeautifulSoup. Se extrajo el título como metadato, se identificaron las secciones (<sec>) descartando aquellas en IGNORE_SECTIONS, y en las secciones válidas se procesaron párrafos (<p>) y tablas (<table-wrap>), transformando cada uno en un registro independiente con número de orden y texto limpio.

Figura 3

Secciones Excluídas del Corpus



Nota. La figura muestra el conjunto de atributos sec-type descartados durante el parseo: references, ref-list, back, ack, acknowledgments, funding, author-notes, app-group, notes y supplementary-material. Su exclusión elimina contenido administrativo y bibliográfico no esencial para el análisis de contenido clínico. Adaptado de. *Autoría propia.*

Normalización y Control de Calidad

Durante la extracción se aplicaron transformaciones automáticas: eliminación de espacios redundantes, consolidación de saltos de línea, filtrado de elementos vacíos y descarte de secciones administrativas según sec-type. El campo order permite reconstruir la secuencia original de cada documento.

Trazabilidad. cada registro mantiene su vínculo con el documento fuente mediante los campos doc_id, article_title, section y order.

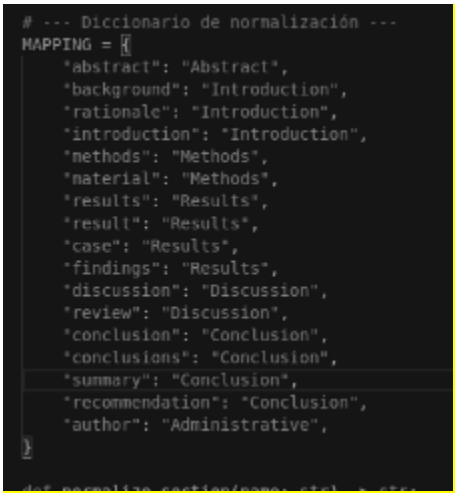
Exploración Estadística de Nomenclatura. Se realizó una exploración estadística para caracterizar la distribución de secciones y detectar inconsistencias en la nomenclatura de las etiquetas <sec>. Este análisis identificó variaciones terminológicas que requerían normalización en fases posteriores.

Los archivos pmc_jats_contenido_limpio.csv presentaban más de treinta variantes terminológicas para las categorías estándar de artículos científicos (Results, RESULTS, 3. Results, etc.). Para resolver esta inconsistencia, se aplicó un proceso de normalización que consolidó las variantes en ocho macro-secciones estandarizadas: Introduction, Methods, Results, Discussion, Conclusion, Administrative, Abstract y Other.

Se diseñó un diccionario de correspondencias que mapeaba las principales variantes terminológicas hacia estas macro-secciones. Cada nombre de sección se convirtió a minúsculas y se comparó con las claves del diccionario. Si existía coincidencia, se asignaba la macro-sección correspondiente; de lo contrario, se clasificaba como Other. Los elementos administrativos como Author contributions se etiquetaron explícitamente para exclusión posterior.

Figura 4

Diccionario de Mapeo de Términos Originales a Macro-Secciones



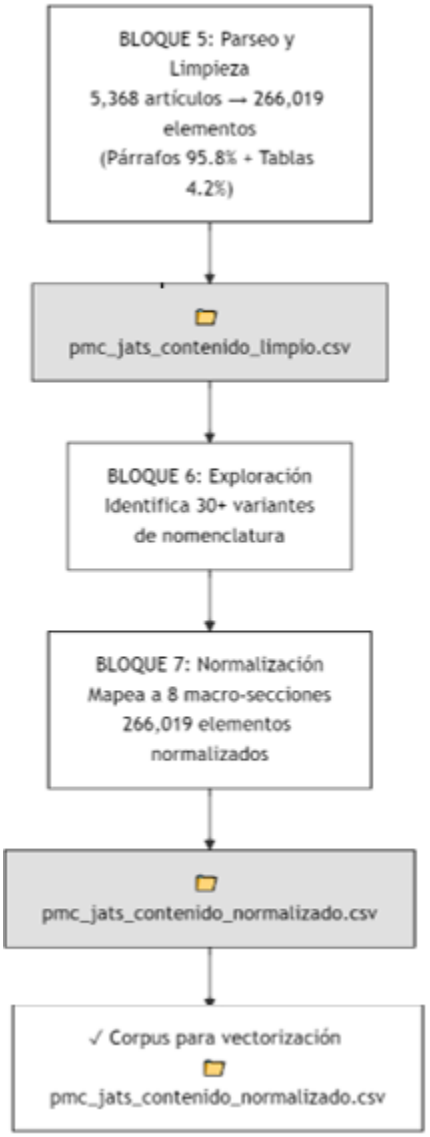
```
# --- Diccionario de normalización ---
MAPPING = {
    "abstract": "Abstract",
    "background": "Introduction",
    "rationale": "Introduction",
    "introduction": "Introduction",
    "methods": "Methods",
    "material": "Methods",
    "results": "Results",
    "result": "Results",
    "case": "Results",
    "findings": "Results",
    "discussion": "Discussion",
    "review": "Discussion",
    "conclusion": "Conclusion",
    "conclusions": "Conclusion",
    "summary": "Conclusion",
    "recommendation": "Conclusion",
    "author": "Administrative",
}
```

Nota. El diccionario implementa equivalencias bidireccionales entre variantes originales y macro-secciones estandarizadas, permitiendo normalizar inconsistencias de capitalización, numeración y redacción presentes en los metadatos XML JATS.

El procedimiento generó el archivo pmc_jats_contenido_normalizado.csv, donde cada elemento del corpus recibió su macro-sección correspondiente. Adaptado de. *Autoría propia.*

Figura 5

Pipeline Minería de Texto. Parseo, Exploración y Normalización



Nota. El diagrama ilustra los tres bloques finales del pipeline de minería de texto. El Bloque 5 transforma 5.368 artículos en 266.019 elementos (párrafos y tablas). El Bloque 6 identifica variantes terminológicas en las etiquetas de sección. El Bloque 7 normaliza estas variantes mediante mapeo a ocho macro-secciones estandarizadas, generando el archivo pmc_jats_contenido_normalizado.csv. Adaptado de. *Autoría propia.*

Implementación de Segmentación Semántica y Vectorización

Se procedió a realizar una segmentación semántica sobre los 266.019 elementos del corpus en pmc_jats_contenido_normalizado.csv. El objetivo fue dividir cada unidad textual en fragmentos coherentes que sirvieran como entrada para vectorización.

Se empleó el algoritmo SemanticChunker de LangChain Experimental, que detecta rupturas temáticas basándose en similitud semántica entre partes consecutivas. A diferencia de métodos tradicionales que fragmentan por longitud fija o puntuación, analiza el significado mediante embeddings vectoriales, convirtiendo cada fragmento en una representación numérica que mide su similitud semántica.

El modelo empleado fue sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, un modelo multilingüe de 384 dimensiones optimizado para comparaciones semánticas en múltiples idiomas.

Se configuró el parámetro breakpoint_threshold_amount = 95, indicando que el algoritmo calcula distancias coseno entre embeddings consecutivos y selecciona como puntos de quiebre aquellos que superan el percentil 95. En términos prácticos, solo el 5% de las mayores diferencias semánticas generan una división.

Este umbral busca equilibrar dos objetivos: evitar fragmentación excesiva preservando coherencia de párrafos clínicos, y detectar transiciones temáticas significativas como cambios de subtópico en resultados o métodos. Este enfoque prioriza coherencia semántica sobre segmentación mecánica (Martin, 2024).

Limpieza y Refinamiento de Chunks Semántico

Los segmentos generados presentaban una distribución de longitudes con extremos problemáticos: fragmentos menores a 50 caracteres y mayores a 1.500 caracteres. Se definieron

tres umbrales de longitud para estandarizar el corpus:

- HARD_MIN (50 caracteres): descarte de fragmentos muy cortos sin contenido informativo
- MIN_LEN (100 caracteres): límite inferior de aceptación
- MAX_LEN (1.500 caracteres): umbral superior

Los fragmentos que excedían MAX_LEN se subdividieron mediante el algoritmo Split_long_text(), que prioriza divisiones naturales por oraciones (delimitadores .!?). Si no se encontraban separadores adecuados, se aplicaba un corte fijo cada 1.000 caracteres (SUBCHUNK_SIZE).

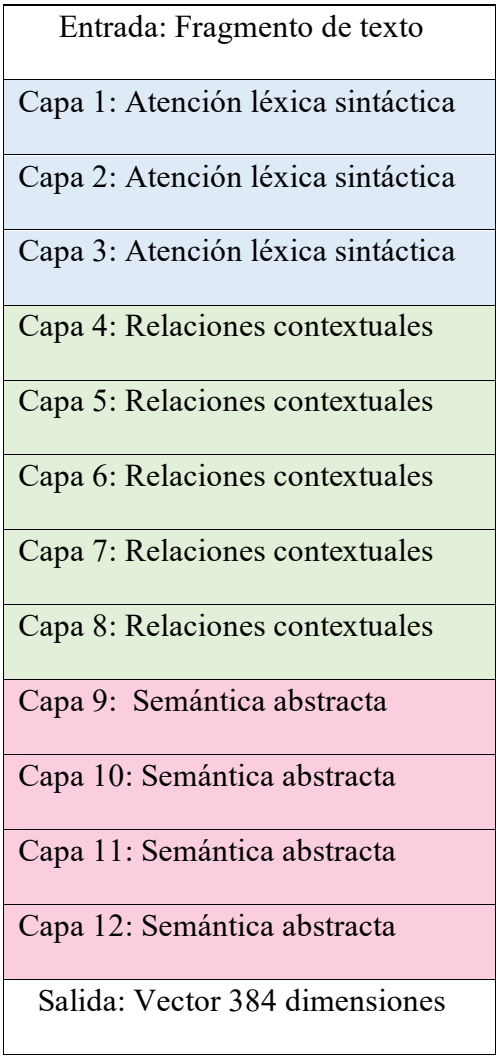
Generación de Embeddings

Se realizó la vectorización de los fragmentos del corpus utilizando el modelo sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, el mismo empleado en la segmentación semántica. Este modelo, desarrollado por Reimers y Gurevych (2019) y distribuido a través de Hugging Face, mapea oraciones y párrafos a un espacio vectorial denso de 384 dimensiones, optimizado para similitud y búsqueda semántica multilingüe.

MiniLM-L12-v2 es un derivado de BERT optimizado para reducir tamaño sin perder rendimiento contextual. Implementa 12 capas de atención y 384 dimensiones ocultas. En una arquitectura Transformer, las capas iniciales capturan patrones léxicos y sintácticos, mientras que las superiores modelan relaciones semánticas más abstractas, generando un vector que sintetiza el significado general del fragmento.

Figura 6

Arquitectura Conceptual del Modelo MiniLM-L12-v2

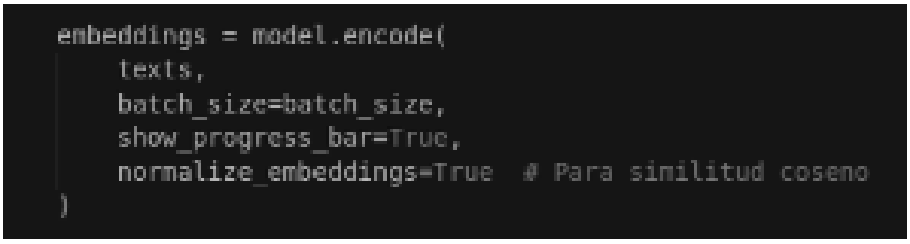


Nota. El diagrama ilustra la estructura jerárquica del modelo: 12 capas de atención donde las iniciales procesan información superficial (léxica-sintáctica) y las posteriores abstraen relaciones semánticas complejas, generando un vector final de 384 dimensiones.

Cada fragmento se convirtió en un vector de 384 componentes numéricos, donde cada dimensión representa una combinación de rasgos lingüísticos y semánticos. Para asegurar comparaciones consistentes, se aplicó normalización L2, ajustando la magnitud de cada vector para que su longitud euclidiana fuera 1.0 mediante el parámetro normalize_embeddings=True. Adaptado de. *Autoría propia.*

Figura 7

Configuración de la Función model.encode para Normalización



```
embeddings = model.encode(
    texts,
    batch_size=batch_size,
    show_progress_bar=True,
    normalize_embeddings=True  # Para similitud coseno
)
```

Nota. La función implementa normalización L2 en los embeddings generados. Adaptado de. *Autoría propia.*

Construcción de Base de Datos Vectorial Persistente

Se utilizó Chroma DB para la indexación de los embeddings generados (como se describió en Metodología). La arquitectura de Chroma incluye el algoritmo Hierarchical Navigable Small World (HNSW) (Malkov & Yashunin, 2018), que realiza búsquedas aproximadas en espacios de alta dimensionalidad organizando los embeddings en una estructura jerárquica de grafos navegables, permitiendo localizar vectores similares de manera eficiente.

El proceso de indexación inició con la carga del archivo pmc_chunks_embeddings_sbert.csv conteniendo los embeddings de vectorización. Dado que los vectores fueron almacenados como cadenas de texto, se aplicó la función ast.literal_eval() para convertirlos en listas numéricas. En Chroma DB se creó una colección persistente configurada con métrica de similitud coseno, almacenada en el directorio local chroma_db_embarazo_sbert_cosine.

La carga se realizó mediante procesamiento por lotes de mil registros, integrando para cada lote el texto del fragmento, su vector asociado, identificador único, y metadatos (identificador del documento, sección normalizada, tipo de elemento, longitud textual).

Cliente Persistente

Se utilizó el cliente persistente de Chroma DB para garantizar continuidad del trabajo sin necesidad de reindexación. Esto permite conservar la base vectorial en disco y acceder a ella en sesiones posteriores, asegurando disponibilidad de metadatos, relaciones entre fragmentos y vectores de similitud sin reconstrucción.

Generación de Pares QA

La generación de pares se realizó localmente mediante Ollama utilizando el modelo gpt-oss:20b, optimizando prompts para generar interrogantes y respuestas clínicas en español.

El sistema operó sobre 57 consultas temáticas predefinidas que abarcaron tópicos generales de salud materna (complicaciones obstétricas, preeclampsia, hemorragia postparto, diabetes) y dominios especializados (microbiota vaginal, anomalías uterinas, farmacología obstétrica, salud neonatal).

Umbrales control de calidad:

- MIN_SIMILITUD_CHUNK_RESPUESTA = 0.65: garantiza alineamiento semántico entre respuesta y fragmento fuente usando embeddings SBERT
- MAX_SIMILITUD_QA_REPETITION = 0.90: evita redundancia excesiva entre pregunta y respuesta
- MIN_LEN_CHUNK_CHARS = 300: filtra fragmentos breves sin contenido informativo suficiente

Durante la ejecución, cada consulta recuperó fragmentos similares mediante el algoritmo HNSW, generando un único par QA. Los pares fueron validados midiendo la similitud chunk-respuesta y disimilitud pregunta-respuesta. Pares que no superaban los umbrales definidos fueron descartados.

Resultados

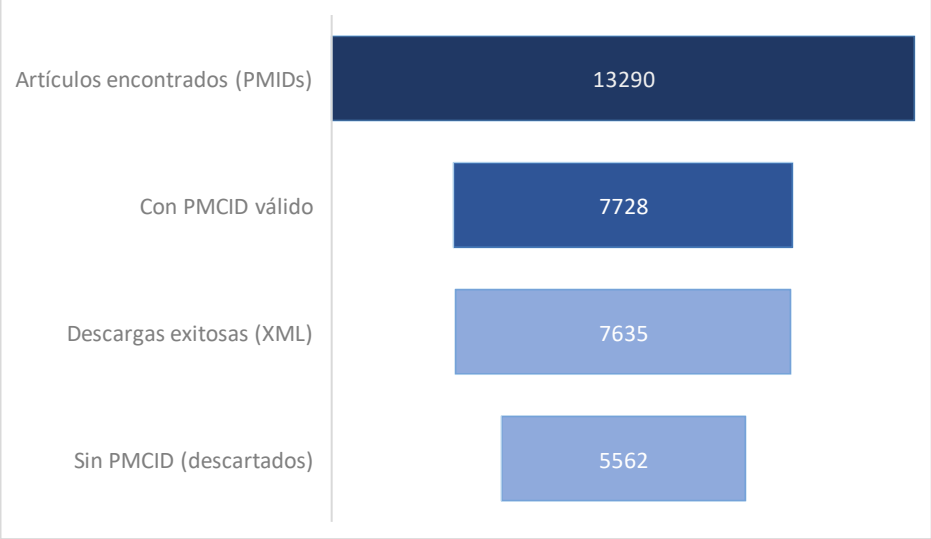
Pipeline Procesamiento y Post – Procesamiento de Datos Biomédicos

Se ejecutó una búsqueda en PubMed con ecuación booleana para literatura en salud materna (2022-2025, acceso abierto). La ecuación combinó: conceptos generales ('maternal health', 'pregnancy'), complicaciones obstétricas ('preeclampsia', 'postpartum hemorrhage', 'gestational diabetes') y restricciones temporales.

La búsqueda retornó 13290 artículos (PMID). El mapeo PMID-PMCID identificó 7728 artículos válidos (58.1%) con acceso a formato XML JATS en PubMed Central. Los 5562 restantes (41.9%) carecían de PMCID y fueron descartados.

Figura 8

Flujo de Recuperación de Artículos Biomédicos Desde PubMed hasta Artículos en Formato XML JATS



Nota. La descarga de archivos XML fue exitosa en 7635 casos (98.8%). Los 93 errores se debieron a fallos de servidor o archivos incompletos. Adaptado de. *Autoría propia.*

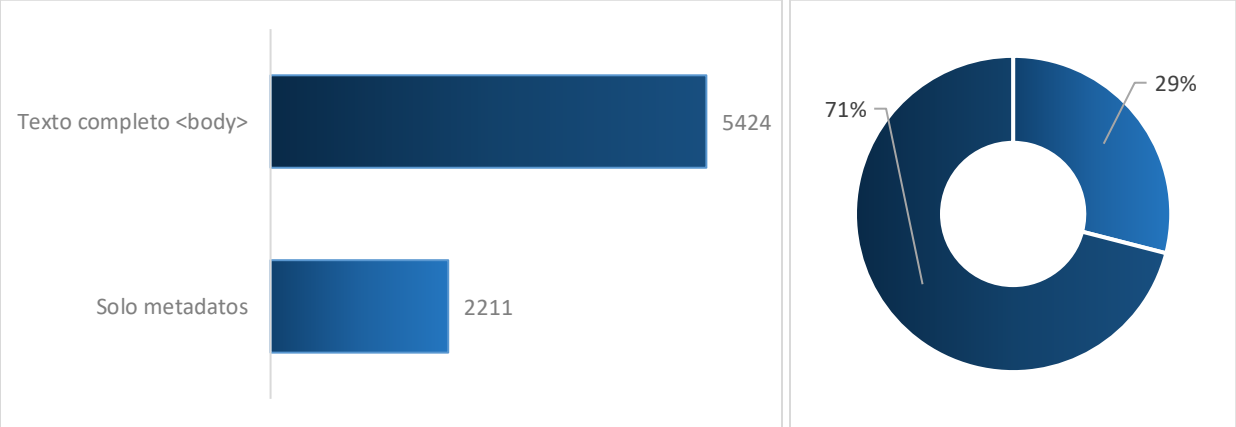
De los 7635 descargas exitosas, se validó la presencia del elemento <body> para diferenciar texto completo de metadatos. Se obtuvieron 5424 artículos con texto completo (71.0%) y 2211

solo con metadatos (29.0%, restricciones editoriales). Solo los 5424 se procesaron en fases posteriores.

El parseo extrajo 266019 elementos: 254802 párrafos (95.8%) y 11.217 tablas (4.2%). Cada elemento conserva vinculación con su documento de origen.

Figura 9

Distribución de Artículos con Texto Completo y Metadatos en Formato XML



Nota. Distribución de artículos con texto completo versus metadatos bibliográficos. Adaptado de. *Autoría propia.*

Tabla 5

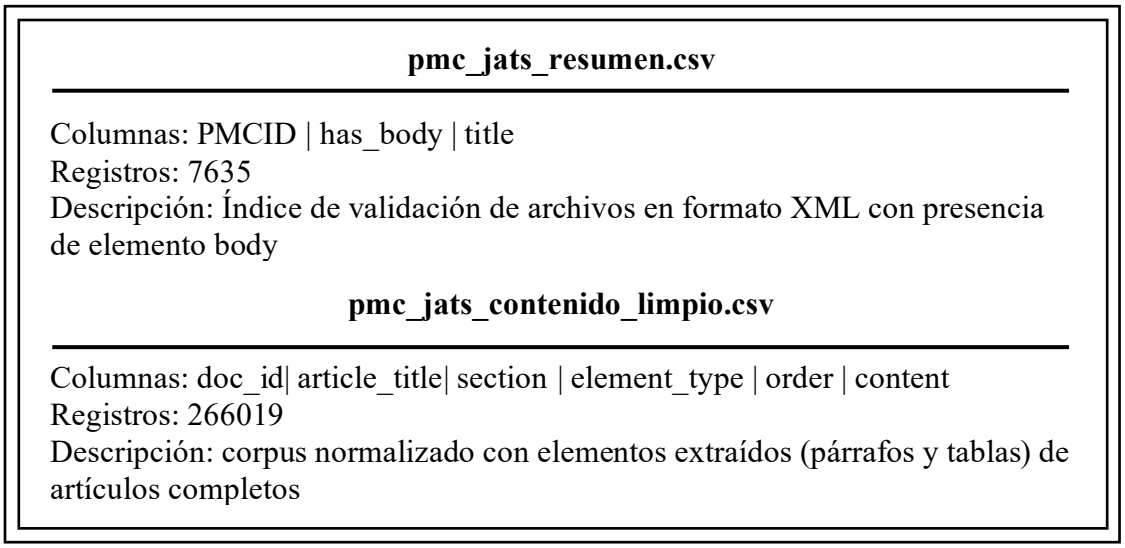
Composición Estructural del Corpus

Elemento	Cantidad	Porcentaje
Párrafos	254802	95.8%
Tablas	11217	4.2%
Total	266019	100%

Nota. Distribución de párrafos y tablas extraídas del corpus XML. Adaptado de. *Autoría propia.*

Figura 10

Archivos CSV Generados Tras Validación Estructural y Extracción de Contenido



Nota. Detalle de los archivos intermedios utilizados en las etapas de normalización y segmentación. Adaptado de. *Autoría propia.*

Normalización y Segmentación de Corpus

La exploración estadística identificó que las secciones estándar (Introduction, Methods, Results, Discussion, Conclusion) aparecían bajo más de 30 variantes: diferencias en capitalización (Results vs RESULTS), numeración (1. Introduction vs 2. Methods), redacción (Material vs Materials and methods) y denominaciones especializadas (Background, Case presentation, Findings).

Tabla 6

Distribución de Variantes de Secciones

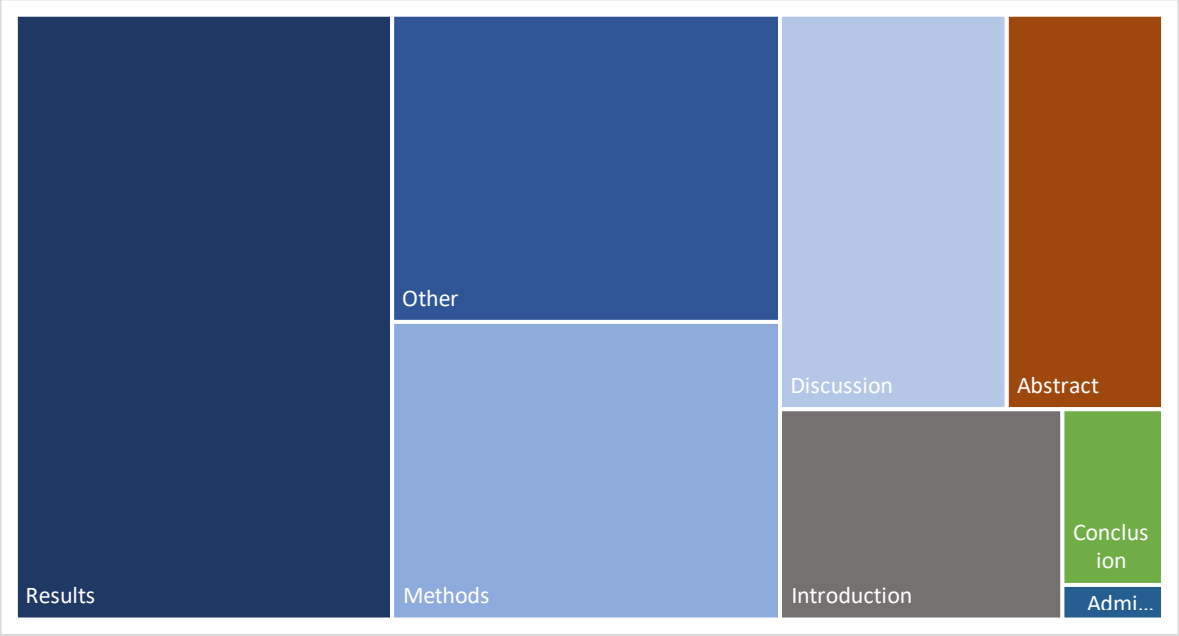
Sección / Encabezado Original	Frecuencia	Macro-sección Asignada
Results	73189	Results
Discussion	27553	Discussion
Methods	24657	Methods
Introduction	12968	Introduction
RESULTS	10431	Results
Materials and methods	5502	Methods

Background	5178	Other
3. Results	4585	Results
4. Discussion	3721	Discussion
1. Introduction	3649	Introduction
Materials and Methods	3577	Methods
DISCUSSION	3406	Discussion
2. Materials and Methods	2895	Methods
METHODS	2102	Methods
INTRODUCTION	1627	Introduction
MATERIALS AND METHODS	1470	Methods
Results and discussion	1224	Results
Review	928	Other
2. Methods	828	Methods
Result	707	Results
Case presentation	700	Other
Material and methods	692	Methods
3. Discussion	655	Discussion
Findings	545	Results
Author contributions	535	Administrative
RATIONALE	526	Other
Recommendations	492	Conclusion
2. Results	486	Results

Nota. Se muestran las 30 variantes más frecuentes y su correspondencia a macro-secciones asignadas. Esta diversidad representa inconsistencias de indexación en metadatos XML JATS. La normalización consolidó estas variantes en 8 macro-secciones: Introduction, Methods, Results, Discussion, Conclusion, Administrative, Abstract, Other. Cada sección se convirtió a minúsculas y se comparó contra un diccionario de correspondencias. Las coincidencias se asignaron a su macro-sección; las demás se clasificaron como Other. Adaptado de. *Autoría propia.*

Figura 11

Distribución Normalizada de Macro-Secciones



Nota. Results concentra el 36.0%, Methods el 18.9%, Other el 18.1%. Las 8 macro-secciones normalizaron completamente el corpus. Adaptado de. *Autoría propia.*

Los hallazgos principales:

Las secciones Introduction, Methods, Results, Discussion y Conclusion agrupan 213301 elementos (79.9%), confirmando prevalencia de estructura IMRaD típica en artículos empíricos. Results (95792 fragmentos, 35.9%) constituye la sección más extensa, reflejando énfasis en hallazgos experimentales y resultados clínicos.

Methods (48469; 18.2%) y Discussion (37.603; 14.1%) mantienen proporciones consistentes con investigación aplicada.

La categoría Other (50245 elementos, 18.8%) agrupa secciones no estandarizadas (Background, Case presentation, Review, Rationale), preservadas para auditorías posteriores.

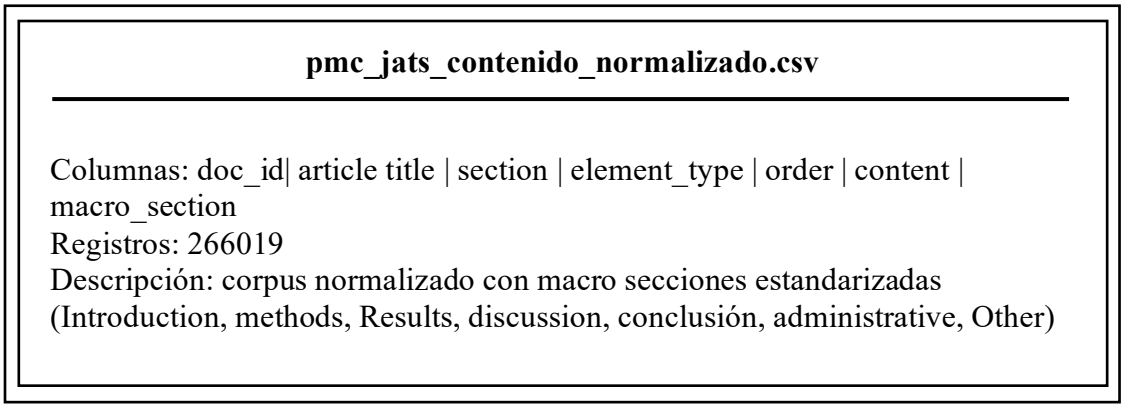
Se identificaron 1447 registros (0.5%) como Administrative (contribuciones de autor, agradecimientos), marcados para exclusión posterior.

Abstract contiene solo 26 registros (0.01%), confirmando eliminación efectiva de resúmenes y enfoque en texto completo.

Los elementos clasificados como Other se conservaron para preservar contenido clínico específico y garantizar integridad del corpus.

Figura 12

Generación del Corpus Normalizado por Macro Sección



Nota. Corpus normalizado con distribución final por macro-sección. Adaptado de. *Autoría propia.*

Segmentación Semántica del Corpus

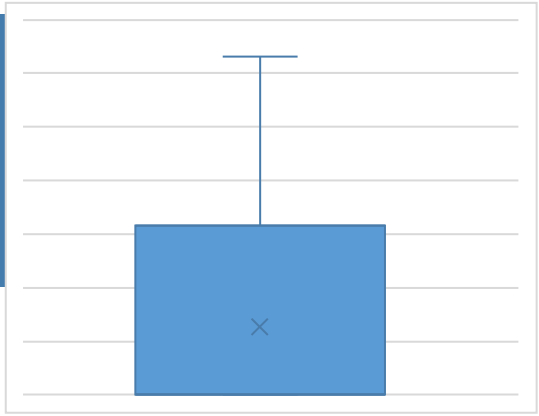
Se aplicó segmentación semántica a los 266019 elementos normalizados utilizando SemanticChunker. Este enfoque detecta rupturas temáticas basándose en similitud coseno entre frases consecutivas, con umbral estricto (percentil 95). A diferencia de fragmentación por longitud fija, prioriza coherencia conceptual: solo transiciones temáticas significativas (5% superior) generan divisiones.

Empleó el modelo MiniLM-L12-v2 para representaciones numéricas. La segmentación generó 417856 chunks (promedio 1.6 por elemento).

Figura 13

Distribución Estadística de la Segmentación Semántica

count	417856.000000
mean	543.804744
std	1491.143772
min	1.000000
25%	143.000000
50%	307.000000
75%	546.000000
max	126175.000000



Nota. Distribución de longitudes de chunks: IQR de 143-546 caracteres, mediana de 307. Desviación estándar de 1491,1 indica variabilidad alta por outliers extremos (mínimo 1 carácter, máximo 126175). Adaptado de. *Autoría propia.*

Los fragmentos breves (<50 caracteres) carecen de contexto clínico. Los fragmentos extensos (>1.500 caracteres), aunque temáticamente coherentes, diluyen precisión del embedding al promediar múltiples ideas, reduciendo especificidad de respuestas.

Se estableció rango óptimo de 300-1500 caracteres. Fragmentos menores al percentil 50 (307 caracteres) contenían predominantemente información no clínica.

Se procesaron los 417856 chunks con dos operaciones: descarte de fragmentos <300 caracteres y re-chunking de fragmentos >1500 caracteres.

Figura 14

Corpus Generado por Segmentación Semántica

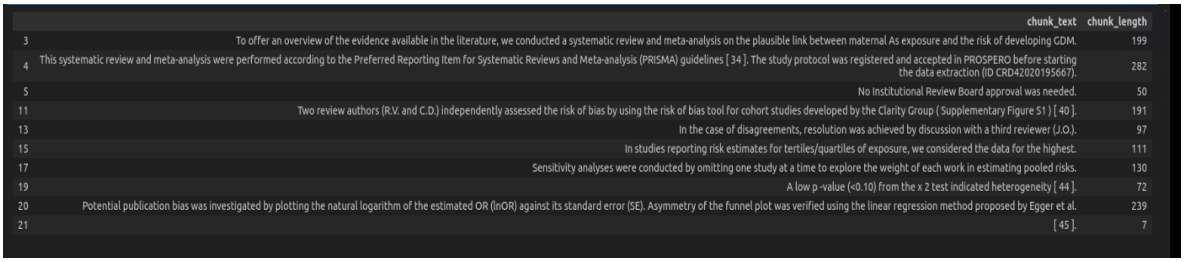
pmc_chunks_semantic.csv

Columnas: doc_id| chunk_id| chunk_text | chunk_length | macro_section | element_type | order
Registros: 417856

Nota. Corpus segmentado semánticamente: 417856 registros con outliers. Requirió limpieza posterior. Adaptado de. *Autoría propia.*

Figura 15

Muestra Aleatoria de Fragmentos Menores al Cuartil 50

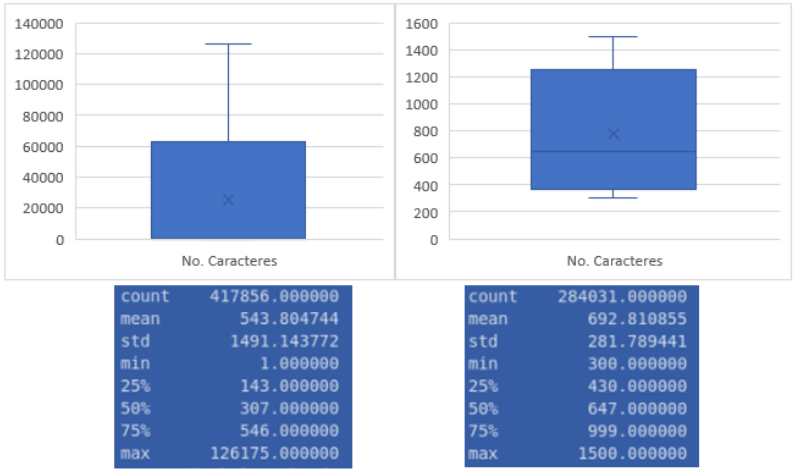


	chunk_text	chunk_length
3	To offer an overview of the evidence available in the literature, we conducted a systematic review and meta-analysis on the plausible link between maternal As exposure and the risk of developing GDM.	199
4	This systematic review and meta-analysis were performed according to the Preferred Reporting Item for Systematic Reviews and Meta-analysis (PRISMA) guidelines [34]. The study protocol was registered and accepted in PROSPERO before starting the data extraction (ID CRD42020195667).	282
5	No Institutional Review Board approval was needed.	50
11	Two review authors (R.V. and C.D.) independently assessed the risk of bias by using the risk of bias tool for cohort studies developed by the Clarity Group (Supplementary Figure S1) [40].	191
13	In the case of disagreements, resolution was achieved by discussion with a third reviewer (J.O.).	97
15	In studies reporting risk estimates for tertiles/quartiles of exposure, we considered the data for the highest.	111
17	Sensitivity analyses were conducted by omitting one study at a time to explore the weight of each work in estimating pooled risks.	130
19	A low p -value (<0.10) from the x 2 test indicated heterogeneity [44].	72
20	Potential publication bias was investigated by plotting the natural logarithm of the estimated OR (lnOR) against its standard error (SE). Asymmetry of the funnel plot was verified using the linear regression method proposed by Egger et al.	239
21	[45].	7

Nota. Muestra aleatoria de fragmentos <307 caracteres (percentil 50). Todos carecen de contenido clínico sustantivo, se excluyen del corpus. Adaptado de. *Autoría propia.*

Figura 16

Comparación de Distribuciones. Izquierda: Resultados con Outliers. Derecha: Normalizado



Nota. Comparación de distribuciones antes y después de limpieza. Reducción del 81,1% en desviación estándar (1491,1 → 281,8), demostrando homogeneización del corpus. Adaptado de. *Autoría propia.*

Tabla 7

Distribución Final por Macro-Sección

Macro-sección	Cantidad
Results	106979
Other	50936
Discussion	46786
Methods	44152
Introduction	28171
Conclusion	6651
Administrative	339
Abstract	17
TOTAL	284031

Nota. Distribución final del corpus tras limpieza y re-chunking. Results contiene 106979 registros (37.6%), seguida por Other (50936; 17.9%), Discussion (46786; 16.5%) y Methods (44152; 15.5%). El corpus homogéneo contiene 284031 chunks, cada uno con longitud entre 300-1500 caracteres. Adaptado de. *Autoría propia.*

Figura 17

Corpus Final para Generación de Embeddings e Indexación

pmc_chunks_semantic_clean.csv

Columnas: doc_id| chunk_id| chunk_text | chunk_length | macro_section | element_type | order
Registros: 284031

Nota. corpus homogéneo con contenido clínico substancial. Chunks <300 fueron descartados.

Adaptado de. *Autoría propia.*

Corpus Vectorizado y Estructura de Base de Datos

Vectorización e Indexación

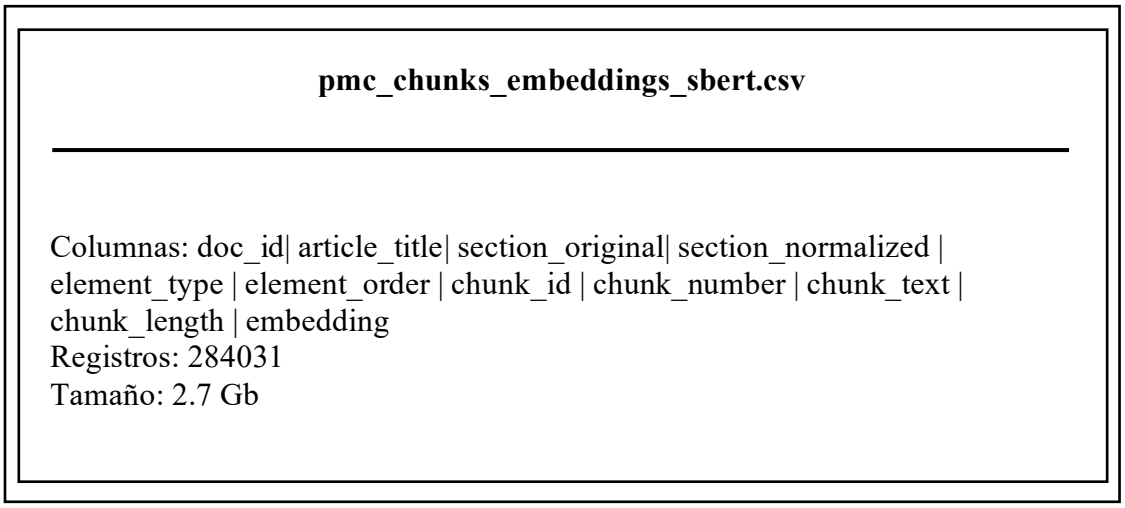
Se vectorizaron los 284031 fragmentos del corpus representando cada uno como vector numérico que captura significado semántico. Esta representación recupera información por similitud conceptual y constituye la base para indexación y posterior generación de pares QA.

El proceso generó una matriz de 284031 × 384 dimensiones donde cada fila representa un fragmento y cada columna una característica semántica. Se utilizó el modelo paraphrase-multilingual-MiniLM-L12-v2 por su capacidad multilingüe, permitiendo capturar significado en español aun cuando la fuente está en inglés.

Durante la vectorización cada fragmento se codificó en espacio de 384 dimensiones y se normalizó mediante norma L2. La normalización ajustó la magnitud de cada vector al rango [-1, 1], permitiendo comparaciones basadas únicamente en significado, sin influencia de la extensión del texto.

Figura 18

Archivo Generado: Pmc_Chunks_Embeddings_Sbert.Csv



Nota. corpus biomédico vectorizado, conserva coherencia semántica y consistencia estructural Ejemplo de embedding procedente del artículo "Maternal Arsenic Exposure and Risk of Gestational Diabetes" (PMCID 7600218), sección Results. Adaptado de. *Autoría propia.*

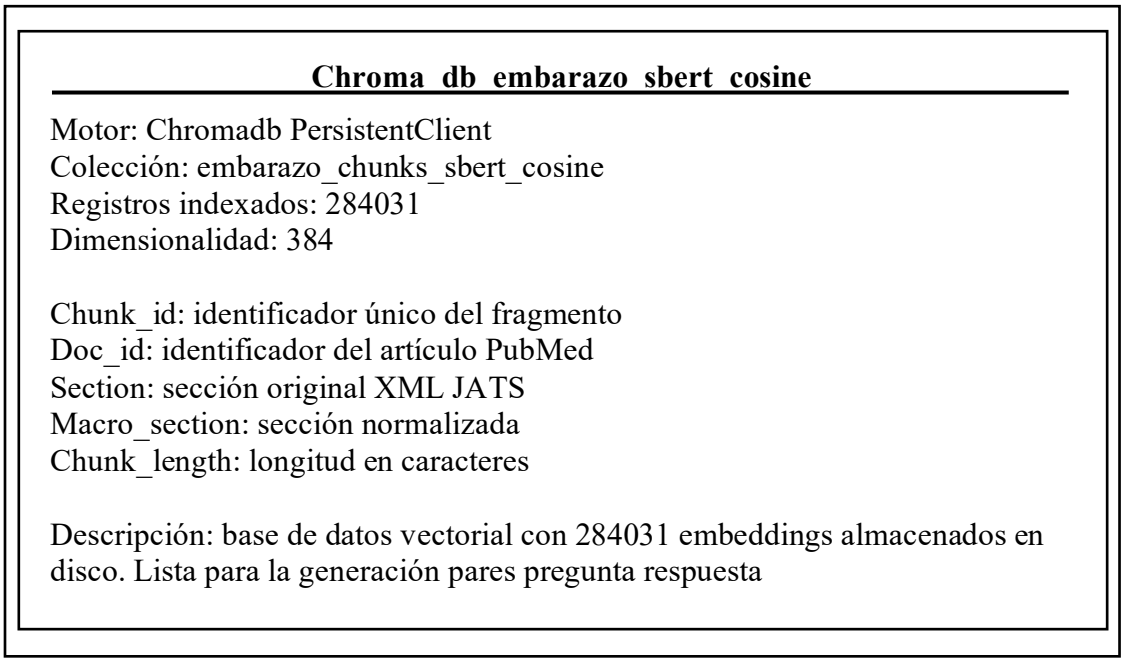
Primeros 10 valores: [-0.0724, 0.0262, -0.0457, 0.0696, 0.0823, 0.0843, 0.0759, 0.0778, 0.0488, -0.0265]

Este vector captura semánticamente el contenido del fragmento (diabetes gestacional, riesgo materno-perinatal, prevalencia en embarazadas). Fragmentos con temas similares generan vectores cercanos en el espacio de 384 dimensiones.

Se indexaron los embeddings en Chroma DB, motor especializado en bases vectoriales. Los 284031 embeddings se incorporaron a la colección embarazo_chunks_sbert_cosine configurada con distancia coseno como métrica de similitud.

Figura 19

Base de Datos Vectorial: Chroma_db_embarazo_sbert_cosine



Nota. Almacenamiento persistente en ChromaDB permite recuperación inmediata de fragmentos por similitud semántica sin recalcular embeddings. Adaptado de. *Autoría propia.*

Corpus Pares QA

Generación y Validación

Se generaron pares QA mediante recuperación semántica vectorial, generación automática con modelo de lenguaje y validación multicriterio. Se definieron 60 consultas estructuradas en cinco categorías clínicas:

- Salud materna base (10 consultas): complicaciones maternas, diabetes gestacional, preeclampsia, hemorragia posparto, infecciones puerperales, mortalidad materna, asfixia neonatal, cuidados prenatales, factores de riesgo, depresión posparto.

- Condiciones específicas (10 consultas): anomalías uterinas, placenta accreta, trombosis, colestasis obstétrica, incompatibilidad Rh, parto prematuro, restricción del crecimiento fetal, eclampsia, síndrome HELLP, ruptura prematura de membranas.
- Microbiota y bacteriología (7 consultas): microbiota vaginal, infecciones bacterianas, Streptococcus agalactiae, candidiasis, tricomoniasis, clamidia, gonorrea.
- Procedimientos clínicos (8 consultas): amniocentesis, ultrasonido prenatal, monitoreo fetal, parto vaginal operatorio, cesárea, episiotomía, inducción de parto, anestesia obstétrica.
- Salud neonatal (7 consultas): apgar score, ictericia neonatal, hipoglucemia neonatal, distress respiratorio, sepsis neonatal, displasia broncopulmonar, enterocolitis necrotizante.
- Psicología y bienestar (6 consultas): ansiedad gestacional, estrés perinatal, violencia doméstica, soporte emocional, bonding materno-fetal, trauma del parto.
- Farmacología y tratamiento (6 consultas): antibióticos en embarazo, anticonvulsivos, antihipertensivos, corticoides prenatales, magnesio sulfato, oxitocina.
- Epidemiología y demografía (6 consultas): disparidades sanitarias, embarazo adolescente, edad materna avanzada, obesidad gestacional, nutrición materna, alcohol y tabaco.

Para cada consulta se recuperaron 10 chunks más similares desde la colección embarazo_chunks_sbert_cosine - Chroma DB. El procedimiento consistió en convertir la consulta a un vector usando el modelo SBERT, luego y por medio de la similitud coseno entre ese vector y los 284031 fragmentos indexados se seleccionaron 10 fragmentos con mayor similitud. El proceso se repitió para cada una de las 60 consultas, generando un conjunto de 600 fragmentos.

El modelo gpt-oss:20b procesó cada fragmento identificando una frase informativa que sea de alto valor clínico, formular la pregunta específica y derecta en español que capture esa información y producir una respuesta que traduzca fielmente la frase original.

Se aplio 3 criterios para garantizar la coherencia semántica y utilidad del par generado

- Longitud mínima: preguntas ≥5 palabras, respuestas ≥10 palabras.
- Similitud chunk-respuesta: ≥0.65 (alineamiento semántico medido con SBERT).
- Disimilitud pregunta-respuesta: ≤0.90 (evita paráfrasis donde pregunta y respuesta son casi idénticas).

Tabla 8

Características de Ejecución Bloque Generación QA

Métrica	Valor
Total de pares generados	100
Tiempo total de ejecución	44:37min
Chunks candidatos procesados	600
Similitud prom (chunk-respuesta)	0.794

Nota. Se generaron 100 pares QA válidos a partir de 600 fragmentos candidatos (60 consultas × 10 fragmentos) mediante validación multicriterio. La similitud chunk-respuesta promedio fue 0.794. Adaptado de. *Autoría propia.*

Evaluación de Calidad Corpus QA

Se evaluaron los 100 pares pregunta-respuesta mediante una métrica de similitud que mide qué tan relacionada está la respuesta generada con el fragmento de texto original. Esta similitud se calculó comparando numéricamente el contenido del fragmento fuente con la respuesta, obteniendo un valor entre 0 y 1, donde 1 representa máxima coherencia y 0 mínima

coherencia. Este valor de similitud permitió clasificar automáticamente cada par en categorías de calidad, proporcionando una evaluación objetiva de cuán bien los pares capturaban información clínica relevante sobre salud materna.

Clasificación por Categorías de Calidad

Se definieron tres categorías de calidad basadas en los valores de similitud, Excelente: similitud ≥ 0.80. Pares donde la respuesta captura fielmente el contenido del fragmento original, sin ambigüedades ni pérdidas de información clínica relevante. Bueno: similitud entre 0.70 y 0.79. Pares coherentes pero con menor precisión, donde la respuesta puede contener información complementaria o tener formulaciones ligeramente distintas a la fuente. Aceptable: similitud entre 0.60 y 0.69. Pares válidos pero con limitaciones evidentes: información parcial, interpretaciones diferentes o coherencia moderada respecto al fragmento original.

Resultados de la Evaluación

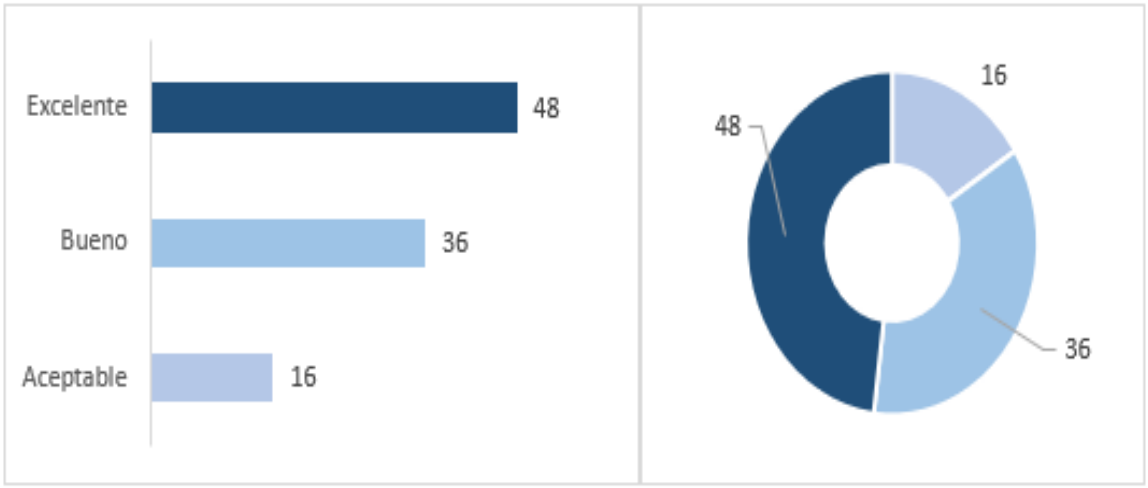
Del total de 100 pares evaluados, la distribución por categorías de calidad fue la siguiente:

- Excelente: 48 pares (48.0%). Casi la mitad del corpus alcanzó la máxima calidad, indicando que el pipeline de generación produjo pares altamente coherentes con los fragmentos originales.
- Bueno: 36 pares (36.0%). Más de un tercio del corpus se clasificó en esta categoría, reflejando coherencia sólida aunque con variabilidad en la precisión de captura de información.
- Aceptable: 16 pares (16.0%). Una minoría de pares presentó limitaciones moderadas pero aun así retuvieron contenido clínico válido.

En conjunto, 84 pares (84.0%) alcanzaron calidad Excelente o Bueno, demostrando que la mayoría del corpus es apto para aplicaciones en sistemas de preguntas y respuestas clínicas.

Figura 20

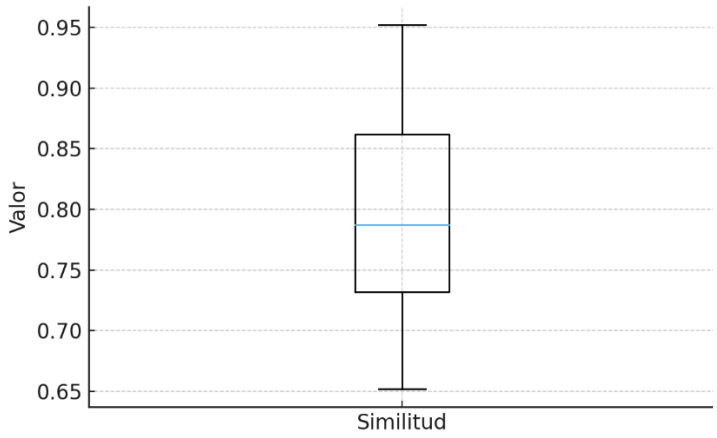
Distribución pares Pregunta-Respuesta Según Categoría de Calidad Semántica



Nota. Los 100 pares fueron clasificados en tres categorías de calidad basadas en similitud coseno: Excelente (≥0.80), Buono (0.70-0.79) y Aceptable (0.60-0.69). El 84% del corpus (48 + 36 pares) alcanzó calidad Excelente o Bueno, demostrando la efectividad del filtrado automático. La gráfica de barras (izquierda) muestra frecuencias absolutas; la gráfica de sectores (derecha) muestra la proporción relativa. Adaptado de. *Autoría propia.*

Figura 21

Distribución Similitud Semantica en el Corpus



Nota. Diagrama de cajas y bigotes que resume las medidas de tendencia central y dispersión de los 100 pares. La línea horizontal dentro de la caja representa la mediana (0.787); los límites de la caja indican el rango intercuartílico (Q1=0.732, Q3=0.862); los bigotes muestran el rango completo (mín=0.652, máx=0.952). La media fue 0.794 (IC 95%: [0.778, 0.811]), reflejando una distribución concentrada en el rango de alta coherencia. Adaptado de. *Autoría propia.*

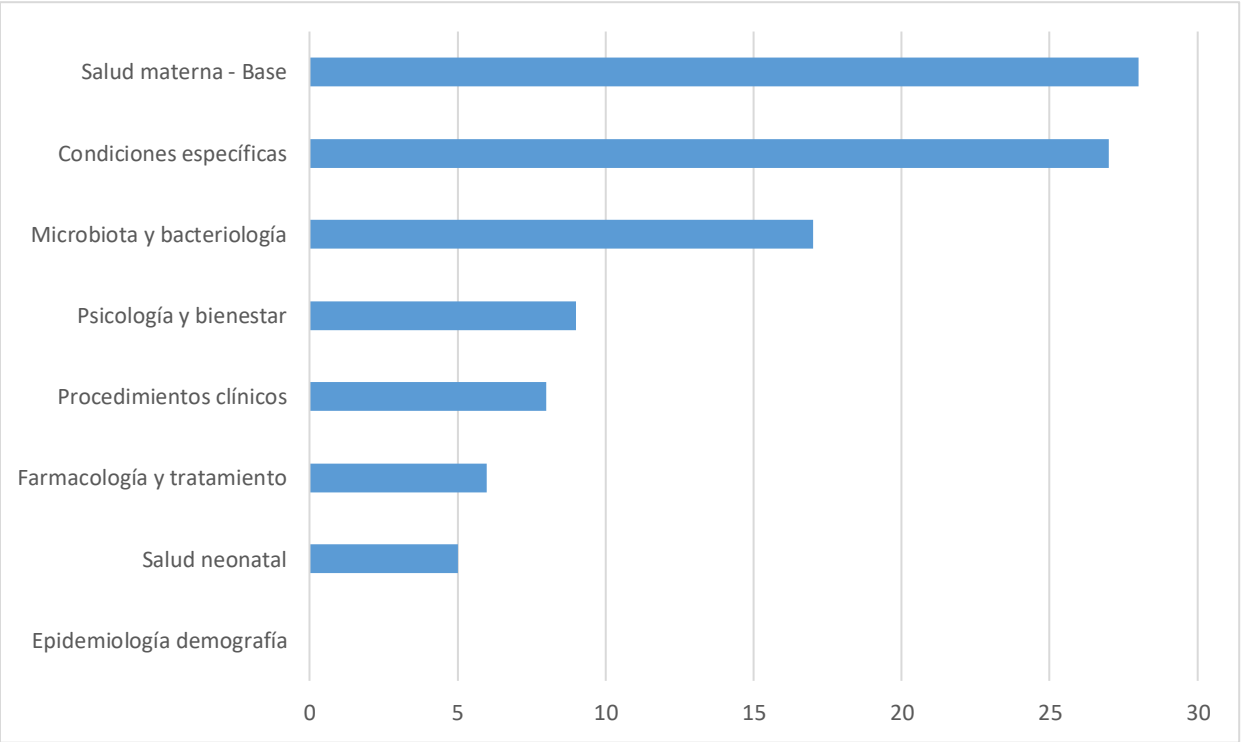
Distribución Temática

Los 100 pares pregunta-respuesta se generaron mediante búsquedas sistemáticas en 8 macrocategorías de salud materna. Sin embargo, solo 7 macrocategorías produjeron pares que cumplieran los criterios de validación automática: Salud materna - Base (28 pares, 28%), Condiciones específicas (27 pares, 27%), Microbiota y bacteriología (17 pares, 17%), Psicología y bienestar (9 pares, 9%), Farmacología y tratamiento (6 pares, 6%), Salud neonatal (5 pares, 5%) y Procedimientos clínicos (8 pares, 8%).

La macrocategoría de Epidemiología y demografía (embarazo adolescente, embarazo edad avanzada, obesidad gestacional, nutrición materna, disparidades sanitarias) no produjo pares validados. Esto se atribuye a que durante la recuperación de fragmentos en la base de datos vectorial, los chunks relacionados con estos temas no alcanzaron el umbral mínimo de similitud semántica (≥0.65) requerido para validar que la respuesta generada fuera congruente con el fragmento original. Adicionalmente, estos temas suelen tratarse en la literatura como contexto epidemiológico o demográfico más que como contenido clínico denso con descripciones detalladas que cumplieran el requisito mínimo de 300 caracteres por fragmento. Esta limitación refleja las características de cobertura de la literatura biomédica disponible en PubMed Central durante el período de búsqueda.

Figura 22

Distribución Pares Pregunta-Respuesta



Nota. Los pares se generaron mediante búsqueda vectorial sistemática y validación automática por similitud semántica coseno. La distribución refleja la disponibilidad de fragmentos biomédicos suficientemente densos y coherentes en PubMed Central. Temas con menor representación (2-3 pares) pueden beneficiarse de expansión manual o búsquedas adicionales especializadas en futuras iteraciones del corpus. Adaptado de. *Autoría propia.*

Conclusiones

Se construyó exitosamente un corpus biomédico de 100 pares pregunta-respuesta en español especializado en salud materna. El pipeline integró nueve fases consecutivas: búsqueda en PubMed (13.290 artículos), descarga XML JATS (5.424 artículos con disponibilidad completa), validación estructural, normalización de secciones (30 variantes → 8 macro-secciones), segmentación semántica (284.031 fragmentos), vectorización con MiniLM-L12-v2, indexación en Chroma DB, generación de pares QA mediante gpt-oss:20b, y evaluación computacional mediante similitud coseno.

Calidad de los Pares Generados

La evaluación mediante similitud coseno identificó tres categorías de calidad semántica: Excelente (similitud ≥0.80) con 48 pares (48%), reflejando respuestas que capturan fielmente el contenido del fragmento original; Bueno (0.70-0.79) con 36 pares (36%), mostrando coherencia sólida con variabilidad moderada en precisión; Aceptable (0.60-0.69) con 16 pares (16%), pares válidos pero con limitaciones evidentes. No se identificaron pares deficientes (similitud < 0.60). La similitud media fue 0.794 (IC 95%: [0.778, 0.811]), indicando distribución concentrada en el rango de alta coherencia semántica. El 84% del corpus alcanzó calidad Excelente o Bueno, demostrando la efectividad de la validación automática.

Contribuciones Metodológicas

Este trabajo aborda tres brechas identificadas en el estado del arte:

Primero, demuestra viabilidad de minería de texto en español para biomedicina, resolviendo la limitante de corpus previos (BioASQ, MedExQA) disponibles solo en inglés.

Segundo, valida segmentación semántica como alternativa a fragmentación por longitud fija, generando chunks limitados por coherencia conceptual mínima en contextos médicos especializados.

Tercero, establece un protocolo de validación bilingüe basado en similitud semántica que evalúa fidelidad de traducciones sin depender exclusivamente de juicio experto.

Operacionalización

Los 284031 embeddings se almacenaron persistentemente en Chroma DB, habilitando búsquedas por similitud semántica en tiempo real. La arquitectura está lista para integración en sistemas RAG y proporciona insumo estructurado para fases posteriores del macroproyecto Minciencias 82244 enfocadas en entrenar modelos de procesamiento de lenguaje natural para telemedicina.

Limitaciones

El tamaño del corpus 100 pares requiere expansión antes de despliegues operacionales. La similitud coseno valida coherencia semántica pero no garantiza corrección clínica: expertos en obstetricia deben revisar pares antes de implementación en sistemas de decisión clínica. La generación automática puede introducir sesgos terminológicos según la disponibilidad de literatura en PubMed Central.

Trabajos Futuros

Validación Clínica

Se requiere someter los 100 pares a revisión por especialistas en obstetricia. Aunque la similitud coseno alcanzó 0.794 con 84% de pares en categorías Excelente o Bueno, esto valida coherencia semántica pero no corrección clínica. Los expertos deben verificar que los pares sean clínicamente válidos y seguros para despliegue en sistemas de telemedicina.

Expansión del Corpus QA

Con 284031 fragmentos disponibles, es viable expandir los pares QA a mínimo 1.000. Se propone paralelizar la generación mediante múltiples instancias del modelo gpt-oss:20b o evaluar modelos alternativos de lenguaje para comparar rendimiento.

Validación Automatizada Mejorada

Para robustecer la evaluación automatizada, se recomienda contrastar similitud coseno (MiniLM-L12-v2) con métricas complementarias como BERTScore, proporcionando múltiples perspectivas sobre alineamiento semántico bilingüe.

Integración en Sistemas RAG

El siguiente paso tecnológico es desplegar el corpus indexado en arquitectura de telemedicina operacional. Se sugiere integrar Chroma DB con modelos de lenguaje avanzados (Claude, GPT-4 vía API) para evaluar rendimiento en escenarios clínicos simulados.

Generalización a Otros Dominios

Se propone adaptar el pipeline a otros dominios de salud críticos en contextos hispanohablantes: pediatría, oncología, enfermedades infecciosas emergentes. Esto validará reproducibilidad metodológica del enfoque más allá de salud materna.

Referencias Bibliográficas

Ankit Pal, L. K. (2022). *MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. https://arxiv.org/abs/2203.14371

Baeseongsu. (2023). *GitHub - baeseongsu/ehrxqa: EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images*. NeurIPS 2023 D&B: https://github.com/baeseongsu/ehrxqa

Boschen, I. (2021). *Software review: The JATSdecoder package—extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed central's open access database*. Scienciometrics : https://doi.org/10.1007/s11192-021-04162-z

datacamp. (28 de 09 de 2024). *Chroma DB Tutorial: A Step-By-Step Guide*. https://www.datacamp.com/tutorial/chromadb-tutorial-step-by-step-guide?utm_cid=21057859163&utm_aid=157296744137&utm_campaign=230119_1-ps-other~dsa~tofu_2-b2c_3-latam-en_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na&utm_loc=9197754-&utm_mtd=-c&utm_kw=&

De Ingeniería del Conocimiento. (2025). *Corpus de calidad: clave para una IA inclusiva y representativa.*

Instituto de Ingeniería del Conocimiento: https://www.iic.uam.es/procesamiento-del-lenguaje-natural/corpus-de-calidad-clave-para-una-ia-inclusiva-y-representativa/

Ekaterina Sviridova, A. Y. (2024). *CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures*.

arXiv:2410.05235

Ekaterina Sviridova, A. Y. (2024). *CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures*. https://arxiv.org/abs/2410.05235

Eyobu, O. N. (2025). *Mother: a maternal online technology for health care dataset.* BMC Res Notes 18, 150: https://doi.org/10.1186/s13104-025-07230-2

Eyobu, O. S. (2025). *Mother: a maternal online technology for health care dataset.* BMC Research Notes, 18(1): https://doi.org/10.1186/s13104-025-07230-2

Ixa-Ehu. (2024). *GitHub - ixa-ehu/antidote-casimedicos.* https://github.com/ixa-ehu/antidote-casimedicos

Jeff Chang, B. C. (2009). *Biopython Tutorial and Cookbook*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://biopython.org/DIST/docs/tutorial/Tutorial-1.48.pdf

Jinhyuk Lee, W. Y. (2020). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. Bioinformatics 1-7: 10.1093/bioinformatics/btz682

Keithhon. (2022). *Paraphrase Multilingual MiniLM L12 V2*. https://dataloop.ai/library/model/keithhon_paraphrase-multilingual-minilm-l12-v2/#:~:text=Meet%20the%20paraphrase%2Dmultilingual%2DMiniLM,Example%20Use%20Cases

Krithara, A. N. (2022). *BioASQ-QA: A manually curated corpus for Biomedical Question Answering [Conjunto de datos]*. En Zenodo (CERN European Organization for Nuclear Research).: https://doi.org/10.5281/zenodo.7655130

Krithara, A. N. (2023). *BioASQ-QA: A manually curated corpus for Biomedical Question Answering.* Scientific Data, 10(1): https://doi.org/10.1038/s41597-023-02068-4

Liu, B., Wei, H., Niu, D., Chen, H., & He, Y. (2020). *Asking Questions the HumanWay: Scalable Question-Answer Generation from Text Corpus*. arXiv:2002.00748v2

Mark Neumann, D. K. (2019). *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing*. arXiv:1902.07669v3

Martin, R. (19 de 09 de 2024). *A Visual Exploration of Semantic Text Chunking*. towardds data science: https://towardsdatascience.com/a-visual-exploration-of-semantic-text-chunking-6bb46f728e30/

microsoft. (2020). *MiniLM: Small and Fast Pre-trained Models for Language Understanding and Generation*. Hugging Face: https://huggingface.co/microsoft/MiniLM-L12-H384-uncased?utm_source=chatgpt.com

Nayak, P. (2024). *Semantic Chunking for RAG*. Medium: https://medium.com/the-ai-forum/semantic-chunking-for-rag-f4733025d5f5

Neha Srikanth, R. S.-G. (2024). *Pregnant Questions: The Importance of Pragmatic Awareness in Maternal Health Question Answering*.

arXiv:2311.09542

Nils Reimers, I. G. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.

e Cornell University: https://doi.org/10.48550/arXiv.1908.10084

Seongsu Bae, D. K.-C. (2023). *EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images*. arXiv:2310.18652

Singhal, K., Azizi, S., Tu, T., & Wei, S. S. (2023). *Large language models encode clinical*. Nature Vol 620: https://doi.org/10.1038/s41586-023-06291-2

Yu. A. Malkov, D. A. (2018). *Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs*. arXiv: https://doi.org/10.48550/arXiv.1603.09320

Yunsoo Kim, J. W. (2024). *MedExQA: Medical Question Answering Benchmark with Multiple Explanations*. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.: 10.18653/v1/2024.bionlp-1.14