ISSN: 1692-7257 - Volumen 2 – Número 48 - 2026

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Estrategia híbrida en cascada para la normalización de direcciones mediante LLMs ajustados para el fortalecimiento del análisis geoespacial en registros de cáncer

A Cascading Hybrid Strategy for Address Normalization using Fine-Tuned LLMs to Strengthen Cancer Registry Geospatial Analysis

Silvio Ricardo Timarán Pereira[1]

Jonathan David Viveros Córdoba[1]

[1] **Universidad de Nariño, Ingeniería de Sistemas, San Juan de Pasto, Nariño, Colombia.**

Recibido: 09 abril 2026. Revisado: 22 junio 2026. Aceptado: 24 julio 2026. Publicado: 28 julio 2026.
Autor de correspondencia: Silvio Ricardo Timarán Pereira

Cómo citar: S. R. Timarán Pereira and J. D. Viveros Córdoba, "Estrategia híbrida en cascada para la normalización de direcciones mediante LLMs ajustados para el fortalecimiento del análisis geoespacial en registros de cáncer", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 48, jul. 2026, art. no. 4512, doi: 10.24054/rcta.v2i48.4512

Esta obra está bajo una licencia internacional Creative Commons Atribución-NoComercial 4.0.

Resumen:

La precisión en el análisis geoespacial del Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP) se ve severamente comprometida por la naturaleza heterogénea y ambigua de las direcciones físicas capturadas de forma manual. Estas inconsistencias en la data de origen limitan el alcance analítico de la infraestructura YACHAY-SIG para la generación de inteligencia territorial multiescalar. En este trabajo, se describe el desarrollo e implementación de un marco metodológico híbrido diseñado para optimizar la integridad de los datos espaciales a través de un flujo de procesamiento en dos fases. La fase inicial despliega un motor de geocodificación determinístico adaptado específicamente a la morfología urbana de la región. Para aquellos registros que presentan una alta complejidad semántica como las nomenclaturas no convencionales de "manzana y lote" se incorpora una segunda etapa basada en Modelos de Lenguaje Extensos (LLM). Este componente integra técnicas de Procesamiento de Lenguaje Natural (NLP), utilizando vectorización TF-IDF y Word2Vec para el discernimiento y corrección de anomalías estructurales en las direcciones. Con el objetivo de maximizar la eficiencia en entornos computacionales de recursos restringidos, se aplicó un protocolo de Ajuste Fino Supervisado (SFT) mediante la técnica de Adaptación de Bajo Rango (LoRA) sobre diversos modelos de código abierto. Las pruebas experimentales revelan que la integración del modelo Llama-3.2-3B-Instruct obtuvo el desempeño superior, alcanzando una precisión de normalización del 99.40%. La validación del sistema ejecutada sobre un conjunto de prueba efectivo de 2,324 direcciones de un corpus total de 15,492 instancias únicas distribuidas en 70% para entrenamiento, 15% para validación y 15% para prueba mediante métricas estándar de clasificación (precisión, exhaustividad y puntuación F1) y la tasa de recuperación geoespacial, evidencia una mejora sustancial en la recuperación espacial, elevándola del 33.02% inicial al 55.13%, logrando mitigar la brecha semántica independientemente de las restricciones cartográficas locales. Esta arquitectura se presenta como un modelo escalable de ingeniería de datos para registros epidemiológicos, proyectando futuras integraciones con servicios como OpenStreetMap y el Instituto Geográfico Agustín Codazzi (IGAC) para fortalecer la cobertura geográfica nacional.

Palabras clave: aprendizaje supervisado, geocodificación, modelos de lenguaje extensos, normalización de datos, procesamiento de lenguaje natural.

Abstract:

La precisión en el análisis geoespacial del Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP) se ve severamente comprometida por la naturaleza heterogénea y ambigua de las direcciones físicas capturadas de forma manual. Estas inconsistencias en la data de origen limitan el alcance analítico de la infraestructura YACHAY-SIG para la generación de inteligencia territorial multiescalar. En este trabajo, se describe el desarrollo e implementación de un marco metodológico híbrido diseñado para optimizar la integridad de los datos espaciales a través de un flujo de procesamiento en dos fases. La fase inicial despliega un motor de geocodificación determinístico adaptado específicamente a la morfología urbana de la región. Para aquellos registros que presentan una alta complejidad semántica como las nomenclaturas no convencionales de "manzana y lote" se incorpora una segunda etapa basada en Modelos de Lenguaje Extensos (LLM). Este componente integra técnicas de Procesamiento de Lenguaje Natural (NLP), utilizando vectorización TF-IDF y Word2Vec para el discernimiento y corrección de anomalías estructurales en las direcciones. Con el objetivo de maximizar la eficiencia en entornos computacionales de recursos restringidos, se aplicó un protocolo de Ajuste Fino Supervisado (SFT) mediante la técnica de Adaptación de Bajo Rango (LoRA) sobre diversos modelos de código abierto. Las pruebas experimentales revelan que la integración del modelo Llama-3.2-3B-Instruct obtuvo el desempeño superior, alcanzando una precisión de normalización del 99.40%. La validación del sistema ejecutada sobre un conjunto de prueba efectivo de 2,324 direcciones de un corpus total de 15,492 instancias únicas distribuidas en 70% para entrenamiento, 15% para validación y 15% para prueba mediante métricas estándar de clasificación (precisión, exhaustividad y puntuación F1) y la tasa de recuperación geoespacial, evidencia una mejora sustancial en la recuperación espacial, elevándola del 33.02% inicial al 55.13%, logrando mitigar la brecha semántica independientemente de las restricciones cartográficas locales. Esta arquitectura se presenta como un modelo escalable de ingeniería de datos para registros epidemiológicos, proyectando futuras integraciones con servicios como OpenStreetMap y el Instituto Geográfico Agustín Codazzi (IGAC) para fortalecer la cobertura geográfica nacional.

Palabras clave: aprendizaje supervisado, geocodificación, modelos de lenguaje extensos, normalización de datos, procesamiento de lenguaje natural.

INTRODUCCIÓN

A nivel global, la oncología pública demanda sistemas de vigilancia epidemiológica con capacidades analíticas que trasciendan el registro de incidencia, priorizando la comprensión de la distribución espacial de la patología. En este escenario, el Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP) constituye el nodo central para el análisis de la carga neoplásica en la región. Sin embargo, la transformación de estos datos en activos estratégicos para la toma de decisiones clínicas y administrativas requiere imperativamente de mecanismos de georreferenciación de alta precisión que habiliten una visualización multiescalar fidedigna.

Como respuesta a esta necesidad técnica, se ha implementado YACHAY-SIG[1], una infraestructura de información geográfica inteligente diseñada para robustecer la capacidad analítica del RPCMP. Esta plataforma se sustenta sobre una arquitectura de datos relacional gestionada en PostgreSQL con el motor geoespacial PostGIS, garantizando una eficiencia óptima en el almacenamiento y procesamiento de variables clínicas y geográficas. El sistema permite la desagregación de la prevalencia e incidencia en unidades territoriales micro-locales (barrios, comunas, corregimientos y veredas), facilitando la síntesis cartográfica mediante mapas de calor y coropletas dinámicas[2].

No obstante, la integridad operativa de YACHAY-SIG está intrínsecamente ligada a la calidad de los datos de origen (input data). La localización de los pacientes se deriva de direcciones postales capturadas en entornos asistenciales, las cuales exhiben una elevada entropía y variabilidad estructural. Diagnósticos previos confirman que la ambigüedad en la nomenclatura obstaculiza el acoplamiento con el Marco Geoestadístico Nacional del Departamento Administrativo Nacional de Estadística (DANE) [3], base cartográfica de referencia oficial [4]. Esta problemática se agudiza en la ciudad de Pasto, donde la coexistencia de nomenclaturas oficiales con descriptores informales basados en “manzanas y lotes” genera una brecha de interoperabilidad semántica significativa [5].

En los últimos años, la evolución de la inteligencia artificial geoespacial (GeoAI) ha transformado la forma en que se procesan y normalizan los datos espaciales no estructurados[6]. Tradicionalmente, los sistemas de información geográfica dependían de reglas rígidas y métodos heurísticos que mostraban limitaciones severas al enfrentar descripciones de ubicación imprecisas, coloquiales o incompletas[7]. Sin embargo, para superar las limitaciones de los enfoques basados en caracteres y esquemas tradicionales, la investigación reciente ha incorporado técnicas de procesamiento de lenguaje natural (PLN) fundamentadas en arquitecturas Transformer[8]. En particular, se ha demostrado que modelos como BERT y RoBERTa mejoran de forma significativa la extracción de entidades geográficas y el emparejamiento de direcciones al capturar la semántica contextual del texto[8],[9].

Más recientemente, la emergencia de los grandes modelos de lenguaje (LLMs) ha introducido un nuevo paradigma en la geocodificación inteligente, permitiendo transformar la predicción de coordenadas en tareas generativas avanzadas y resolver descripciones de localidades complejas con relaciones espaciales relativas[10]. Estas tecnologías abren una vía fundamental para cerrar la brecha semántica en dominios complejos como los registros sociosanitarios y poblacionales[6],[10].

Históricamente, la geocodificación en entornos de nomenclatura no estandarizada ha sido abordada mediante diversas metodologías. A nivel regional, se han desarrollado motores determinísticos basados en algoritmos de concordancia de cadenas (string matching) optimizados para la malla vial local[5]. Aunque estos sistemas proporcionan una base sólida, su rendimiento decrece exponencialmente ante registros no estructurados, frecuentes en áreas periféricas y rurales. Paralelamente, la literatura científica internacional indica que el tratamiento manual de estas coordenadas resulta económicamente inviable ante grandes volúmenes de datos[11], lo que ha catalizado la adopción de modelos de aprendizaje profundo y arquitecturas Transformer (como BERT y RoBERTa) para la resolución de ambigüedades textuales [12], [13].

En la frontera tecnológica actual, la Inteligencia Artificial ha convergido hacia los Modelos de Lenguaje de Gran Escala (LLM), los cuales demuestran capacidades disruptivas en el Procesamiento de Lenguaje Natural (NLP)[14]. Sin embargo, la especialización de modelos fundacionales para tareas geográficas críticas requiere protocolos de adaptación específicos. En este contexto, la técnica de Adaptación de Bajo Rango (LoRA) se ha consolidado como el estándar para el ajuste fino (fine-tuning) de estos modelos, permitiendo que arquitecturas de vanguardia, como Gemma 3 y Llama 3.2, alcancen niveles de precisión superiores en la normalización de direcciones[15].

A pesar de los avances metodológicos descritos, persiste una brecha científica significativa en la normalización automatizada de direcciones con alta entropía[16]. Los enfoques tradicionales, fundamentados en algoritmos determinísticos y reglas heurísticas de concordancia de cadenas, operan bajo el supuesto de una nomenclatura urbana rígida, exhibiendo un colapso operativo ante errores tipográficos severos, abreviaturas informales o la ausencia de componentes normativos esenciales[17]. Con la adopción de las arquitecturas Transformer clásicas, la investigación previa logró mitigar parcialmente este problema mediante la extracción de entidades con base en la semántica contextual[8],[9]; sin embargo, estas arquitecturas se ven severamente limitadas por su naturaleza discriminativa, lo que les impide resolver tareas generativas complejas o interpretar relaciones espaciales relativas expresadas en lenguaje coloquial (por ejemplo, descripciones del tipo “detrás del parque principal, lote

baldío")[18].

La estrategia basada en LLMs aborda esta brecha mediante un cambio de paradigma al transformar la normalización en una tarea generativa estructurada capaz de razonar sobre la ambigüedad espacial [10]. No obstante, la literatura actual carece de marcos híbridos eficientes que integren la precisión determinística de los sistemas de información geográfica tradicionales con la flexibilidad semántica de los modelos de lenguaje ajustados mediante técnicas de bajo costo computacional (LoRA) [13], una carencia fundamental que el presente trabajo busca resolver.

Frente a otras propuestas híbridas reportadas en la literatura que por lo general se limitan a conectar de manera secuencial servicios de geocodificación comercial de caja negra con heurísticas genéricas de limpieza de texto[19], el aporte diferenciador de la presente investigación radica en el diseño de un marco de acoplamiento estrecho (tightly-coupled) optimizado para contextos regionales de recursos limitados. En lugar de depender de modelos masivos privativos, la estrategia propuesta integra un motor determinístico local en PostgreSQL/PostGIS con un agente de lenguaje de código abierto (open-weight) especializado mediante adaptación eficiente de parámetros (LoRA)[20]. Este enfoque no solo garantiza la soberanía y privacidad de los datos epidemiológicos sensibles del Registro Poblacional de Cáncer, sino que opera con alta precisión sobre nomenclaturas informales complejas ("manzanas y lotes") características de ciudades intermedias latinoamericanas, logrando un balance óptimo entre coste computacional, interpretabilidad espacial y tasa de recuperación de coordenadas[19],[20].

El presente artículo expone la implementación de una estrategia híbrida que amalgama la robustez de un motor de geocodificación determinístico con la flexibilidad semántica de un agente de NLP potenciado por LLMs optimizados. Esta arquitectura fortalece la integridad de la vigilancia epidemiológica, proporcionando una base científica sólida para el análisis oncológico con enfoque territorial.

MATERIALES Y MÉTODOS

Materiales

El diseño e implementación del módulo de normalización de direcciones, así como su posterior incorporación en la plataforma YACHAY-SIG, se estructuró a partir de tres componentes esenciales: repositorios de datos oficiales, modelos de lenguaje de gran tamaño (LLM) de última generación y una infraestructura de cómputo de alta capacidad dedicada para el procesamiento.

Fuentes de Datos y Referentes Cartográficos

Como insumo primordial, se utilizaron los registros históricos del Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP), los cuales consolidan los datos clínicos y sociodemográficos de la población diagnosticada en la región.

Para garantizar la reproducibilidad del estudio, el conjunto de datos inicial provisto por el RPCMP constaba de un total de 15.645 registros geográficos crudos de pacientes oncológicos. Durante la fase inicial de auditoría y depuración de datos, se aplicaron filtros rigurosos de control de calidad. Específicamente, se excluyeron 153 registros correspondientes a entradas con coordenadas geográficas nulas o corruptas fuera del perímetro urbano (n = 68), duplicados exactos generados por múltiples ingresos asistenciales de un mismo paciente (n = 45), y descripciones postales con una entropía extrema o vacías de información útil (n = 40). Tras este proceso de depuración sistemática, se obtuvo un corpus curado de 15.492 registros válidos, los cuales fueron particionados de manera estratificada para los subconjuntos experimentales de entrenamiento, validación y prueba (véase la Tabla 1).

Tabla 1: Partición del Conjunto de datos del RPCMP

Categoría / Subconjunto	No. Reg.	Criterios Exclusión /inclusión	Composición / Distribución
Registros Crudos Iniciales	15,645	Extracción total de la base del RPCMP.	Totalidad de notificaciones oncológicas brutas
Registros Excluidos (Depuración)	(153)	Coordenadas nulas o fuera del municipio (n=68). Duplicados exactos (n=45). Entropía extrema / campos vacíos (n=40)	Registros descartados por inviabilidad de geocodificación
Dataset Curado Final	15,492	Registros aptos para procesamiento y normalización.	100% de los datos limpios integrados al flujo híbrido
Subconjunto de Entrenamiento	10,844 (70%)	Asignado para el ajuste fino (fine-tuning) mediante LoRA.	Distribución estratificada por comunas y gravedad
Subconjunto de Validación	2,324 (15%)	Utilizado para el monitoreo de hiperparámetros y early stopping.	Muestra representativa independiente.
Subconjunto de Prueba (Test)	2,324 (15%)	Reservado para la evaluación ciega del rendimiento final.	Casos complejos y rurales de alta entropía.

Fuente: Elaboración propia

Para asegurar la precisión en la localización espacial y el proceso de geocodificación, se articularon dos instrumentos cartográficos clave:

Marco Geoestadístico Nacional (MGN): Suministrado por el DANE[3], este recurso permitió establecer la estructura político-administrativa oficial, abarcando desde la división por barrios y sectores urbanos hasta la organización rural en corregimientos y veredas.

Motor de Geocodificación de Pasto: Se integró la herramienta de código abierto desarrollada por Timarán et al.[5], la cual

emplea algoritmos de concordancia de cadenas (string matching) para convertir direcciones físicas en coordenadas geográficas. Este motor resulta esencial al estar calibrado específicamente para la topología vial y la nomenclatura propia de San Juan de Pasto.

Modelos de Lenguaje de Gran Escala (LLMs)

Con el fin de realizar las fases de experimentación y ajuste fino (fine-tuning), se seleccionaron cinco variantes pertenecientes a dos familias de modelos con pesos abiertos (Open Weights). La elección de estos modelos respondió a la necesidad de equilibrar la potencia de procesamiento con la eficiencia operativa en arquitecturas de hardware con capacidades restringidas:
Serie Gemma 3 IT (270M, 1B y 4B): Desarrollados por Gemma Team[21], estos modelos representan el estado del arte en arquitecturas ligeras. Se caracterizan por optimizar el razonamiento lógico y el cumplimiento de instrucciones complejas, manteniendo un consumo energético y de memoria reducido.

Serie Llama 3.2 (1B-Instruct y 3B-Instruct): Se implementaron estas versiones compactas de Meta AI[15], las cuales han sido refinadas para despliegues de alto rendimiento. Destacan por su avanzada capacidad de comprensión semántica y por garantizar una baja latencia en entornos computacionales limitados.

Ecosistema de software y herramientas de desarrollo

Con el propósito de garantizar la replicabilidad científica y promover la soberanía tecnológica, la infraestructura de software se estructuró en su totalidad a partir de herramientas de código abierto. Los componentes del ecosistema se detallan a continuación:

Gestión de Datos Espaciales: Se adoptó PostgreSQL como sistema de gestión de bases de datos relacionales, el cual se complementó con la extensión PostGIS para la administración, almacenamiento y ejecución de consultas de geometrías referenciadas en el sistema de coordenadas WGS84.

Entorno de Programación y Dependencias: El desarrollo de la lógica del sistema se realizó en Python[22], empleando el gestor de paquetes uv para garantizar una administración eficiente y aislada de las dependencias.

Entrenamiento y Ajuste Fino (Fine-Tuning): Las fases de entrenamiento y optimización de los modelos se ejecutaron mediante el uso combinado de las herramientas LLaMA Factory [23] y Unsloth. La adaptación de los modelos se fundamentó en técnicas de parametrización eficiente LoRA (Low-Rank Adaptation)[24], ejecutadas sobre infraestructura gráfica de alto rendimiento con soporte CUDA.

Visualización de Resultados: La representación espacial de la información se integró en la interfaz web de YACHAY-SIG[1], una plataforma desarrollada sobre arquitecturas web modernas orientada a la generación y análisis visual de mapas de calor dinámicos y mapas de coropletas[2].

Infraestructura de Hardware y Despliegue

El proceso de entrenamiento y ajuste fino de los modelos se sustentó en una estación de trabajo de alto rendimiento equipada con una unidad de procesamiento gráfico (GPU) NVIDIA A100 Tensor Core (con 40 GB de VRAM) compatible con la tecnología CUDA. El flujo computacional se desarrolló sobre el sistema operativo Ubuntu 22.04 LTS, utilizando Python 3.10.12, PyTorch 2.3.0 [22] y la biblioteca Transformers de Hugging Face[25], empleando LLaMA-Factory como interfaz principal[23].

Posteriormente, para garantizar la viabilidad técnica y la eficiencia en la etapa de despliegue e inferencia en producción, los modelos optimizados fueron convertidos y cuantizados al formato GGUF (Q5_K_M)[26]. Esta estrategia permitió una gestión óptima de los recursos computacionales y una reducción sustancial en la latencia de respuesta, sin comprometer la precisión espacial obtenida durante el entrenamiento.

Métodos

El marco metodológico de este estudio se fundamenta en un enfoque de investigación aplicada, el cual fue organizado mediante una serie de etapas secuenciales y articuladas. Esta estructura permite transitar de manera rigurosa desde el diagnóstico y tratamiento de las fuentes de datos primarios hasta la validación funcional del sistema inteligente una vez integrado en la arquitectura global.

Fase de Diagnóstico y Caracterización de Dato

Durante la etapa inicial, se llevó a cabo un examen detallado de la base de datos correspondiente al Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP). Mediante la implementación de técnicas de Procesamiento de Lenguaje Natural (NLP), se analizaron sistemáticamente las inconsistencias en los registros, detectando patrones críticos de error como el empleo de abreviaturas no convencionales, la omisión de conectores gramaticales y la recurrencia de sistemas de dirección

basados en esquemas de manzanas y lotes[4].

Este análisis exploratorio fue fundamental para definir los parámetros y requerimientos específicos de entrenamiento, orientando así el ajuste de los modelos de lenguaje hacia la resolución de las particularidades léxicas y estructurales identificadas en la data local.

Construcción del Corpus para el Ajuste Fino

Con el objetivo de especializar el modelo de lenguaje (LLM), se desarrolló un conjunto de datos de entrenamiento (dataset) estructurado a través de dos procesos fundamentales:

Etiquetado Manual (Ground Truth): Se realizó un procedimiento de normalización exhaustiva donde expertos en el dominio analizaron y estandarizaron manualmente cada registro conforme a la normativa oficial de direcciones. Este proceso permitió establecer la "verdad fundamental" o ground truth, indispensable para garantizar la calidad del aprendizaje del modelo.

Formateo de Instrucciones (Instruction Tuning): Los datos fueron organizados bajo un esquema de instrucción diseñado para el aprendizaje supervisado. Esta etapa consistió en la creación de pares lógicos de Entrada (dirección con ruido o inconsistencias) -> Salida (dirección normalizada y clasificación estructural). Dicha estructura facilitó la preparación de los datos para la fase de Entrenamiento Supervisado Continuo (SFT), optimizando la capacidad del modelo para interpretar y corregir nomenclaturas urbanas complejas.

Configuración del Ajuste Fino (Fine-Tuning) mediante LoRA

El componente central de la metodología radicó en la implementación de la técnica de Adaptación de Bajo Rango (LoRA) [24]. En lugar de ejecutar un reentrenamiento integral de los parámetros en los modelos Gemma 3 y Llama 3.2, se optó por la inserción de matrices de bajo rango en la totalidad de sus capas lineales. Este procedimiento técnico ofreció tres ventajas fundamentales:

Optimización de Recursos: Permitió una reducción drástica en el consumo de memoria VRAM al congelar los pesos originales y actualizar exclusivamente los parámetros de adaptación.

Preservación de Capacidades: Facilitó la especialización del modelo en la semántica y nomenclatura local de la ciudad de Pasto, evitando la degradación de las capacidades de razonamiento general inherentes al LLM base.

Eficiencia Computacional: El entrenamiento se realizó en formato de precisión mixta bfloat16 (bf16), optimizado mediante el uso de Gradient Checkpointing y Flash Attention.

A nivel operativo, el proceso de entrenamiento se configuró para un total de 3 épocas con un tamaño de lote efectivo (batch size) de 16 (utilizando un tamaño de lote por dispositivo de 4 y una acumulación de gradientes de 4 pasos) para optimizar la estabilidad en la actualización de los pesos. Se implementó una tasa de aprendizaje inicial de 2×10^{-4} con un planificador de tipo coseno (cosine learning rate scheduler) y un porcentaje de calentamiento (warmup ratio) del 0.03[27].

La estrategia de validación evaluó el desempeño del modelo al final de cada época utilizando el subconjunto de validación, aplicando un criterio de parada temprana (early stopping) basado en la minimización de la pérdida de validación (validation loss) con un margen de paciencia de 2 épocas para prevenir el sobreajuste (overfitting)[28]. En cuanto a la parametrización específica de LoRA, se estableció un factor de rango (r) de 16, un factor de escalado (α) de 32 y una tasa de dropout de 0.05 aplicada sobre las proyecciones de atención[23].

Implementación del Algoritmo Híbrido de Geocodificación

La transformación de descripciones textuales en coordenadas espaciales se fundamentó en un flujo de trabajo bidireccional y secuencial, diseñado para maximizar la tasa de recuperación de registros.

Fase 0, Preprocesamiento y limpieza textual

Antes de la evaluación geoespacial, las cadenas de texto originales se someten a un módulo de limpieza estructurado. Mediante el uso de expresiones regulares, el algoritmo estandariza el uso de mayúsculas, elimina puntuación anómala y ejecuta una separación sistemática de caracteres alfabéticos y numéricos concatenados (por ejemplo, convirtiendo "24A" en "24 A"). Esta etapa preliminar es crítica para mitigar el ruido semántico antes de cualquier proceso de normalización o búsqueda.

Fase 1, Motor determinístico mejorado (búsqueda por proximidad)

En esta etapa, el proceso emplea un geocodificador base[5], el cual fue robustecido mediante heurísticas avanzadas y una mayor flexibilidad sintáctica. A diferencia del enfoque original de orden estricto, el nuevo algoritmo permite interpretar nomenclaturas invertidas o parciales. Además, se sustituyó el criterio de coincidencia exacta por un sistema de búsqueda difusa (fuzzy matching) con una escala de confianza de 0 a 1.

Bajo este esquema, si una dirección no se localiza de forma literal en la malla vial, el sistema evalúa ubicaciones adyacentes aplicando coeficientes de penalización por distancias numéricas o variaciones en sufijos alfabéticos. Esto permite asignar la coordenada de la ubicación más cercana posible, siempre que se cumpla con un umbral de confianza métrica predefinido.

Fase 2, Recuperación inteligente mediante LLM

Como última instancia de procesamiento, aquellos registros que presentan ambigüedades severas, referencias coloquiales extremas o que resultan en un fallo total del motor determinístico, son derivados al modelo de lenguaje (LLM) previamente ajustado. En esta fase, el LLM opera como un traductor semántico capaz de inferir la intención geográfica del texto y reconstruir la dirección bajo los estándares oficiales del DANE.

Una vez normalizada, la cadena es reinyectada en la Fase 1. Este segundo intento permite que el motor determinístico procese una estructura predecible, logrando la georreferenciación exitosa en el sistema WGS84 de datos que, bajo métodos convencionales, serían descartados por su baja calidad.

Protocolo de Evaluación y Métricas de Éxito

Con el objetivo de validar el rendimiento del sistema híbrido, se estableció un protocolo de evaluación fundamentado en dos dimensiones complementarias: la precisión lingüística de la normalización y la efectividad operativa de la geocodificación.

Evaluación lingüística (calidad de la normalización)

Para determinar la solvencia técnica de los modelos ajustados (Gemma 3 y Llama 3.2), se utilizó la métrica de exactitud (accuracy). El cálculo se realizó mediante el contraste directo entre las direcciones normalizadas por el modelo en el conjunto de prueba y la verdad fundamental (ground truth) establecida mediante el etiquetado manual experto. Esta evaluación permitió medir la capacidad de los modelos para abstraer el ruido textual, corregir errores ortográficos y reestructurar la información bajo los estándares de nomenclatura oficial definidos por el DANE[3].

Evaluación de efectividad (tasa de recuperación)

Debido a que la precisión espacial definitiva está condicionada por la cobertura cartográfica y el motor determinístico, la métrica principal del sistema híbrido se definió como la tasa de recuperación de geocodificación (Geocoding Match Rate).

Este indicador representa el porcentaje de direcciones o registros geográficos que el sistema logra ubicar y asociar correctamente con sus coordenadas espaciales o puntos de referencia correspondientes dentro de la base de datos geográfica oficial. Operacionalmente, un valor alto indica que el modelo es capaz de procesar y estructurar descripciones de ubicación imprecisas o incompletas de los ciudadanos, traduciéndolas de manera exitosa en mapas digitales sin pérdida significativa de información.

Esta métrica cuantifica el impacto operativo del pipeline al medir la proporción de registros médicos que, habiendo fallado inicialmente en el motor determinístico tradicional, logran obtener una coordenada válida tras ser procesados por el flujo híbrido (normalización semántica con LLM y búsqueda difusa). Por consiguiente, este indicador no solo evalúa la capacidad interpretativa del modelo, sino el incremento neto en la cobertura geoespacial del sistema, reconociendo la interdependencia entre la mejora semántica y la infraestructura cartográfica disponible.

RESULTADOS Y DISCUSIÓN

El análisis de la estrategia implementada permitió corroborar la efectividad de los modelos de lenguaje de gran escala (LLM) como herramientas para la recuperación de datos geoespaciales esenciales para el RPCMP. A continuación, se detallan los hallazgos estructurados en el diagnóstico de calidad, el desempeño del proceso de ajuste fino y el impacto obtenido en la georreferenciación definitiva.

Diagnóstico de la Calidad de los Datos en el RPCMP

Con el fin de establecer una línea base y dimensionar la complejidad técnica de los registros de cáncer en el municipio de Pasto, se efectuó un diagnóstico inicial empleando el geocodificador de código abierto de Timarán et al.[5]. Este sistema, aunque representa un pilar tecnológico regional, posee una arquitectura determinista diseñada para estructuras viales predecibles, lo que limita su desempeño ante datos con alta variabilidad.

El diagnóstico ratificó que las inconsistencias y ambigüedades presentes en el RPCMP constituyen una barrera crítica para los sistemas basados en reglas rígidas. Las direcciones capturadas en contextos clínicos suelen presentar:

- Errores tipográficos y omisiones gramaticales.
- Abreviaturas no estandarizadas.
- Descripciones coloquiales ajenas a los algoritmos convencionales de concordancia de cadenas (string matching).

Un factor metodológico clave fue el procesamiento de los registros bajo los dos modelos de nomenclatura que coexisten en la ciudad: “malla vial” y “barrio-manzana”. Se aplicó un protocolo exhaustivo donde cada dato fue evaluado por ambas funciones, validando el resultado si cualquiera de ellas lograba la asignación de coordenadas.

Los resultados sobre una muestra de 15,645 registros evidenciaron la magnitud del desafío: únicamente 5,167 direcciones (33.02 %) fueron geocodificadas exitosamente. En contraste, 10,478 registros (66.97 %) no pudieron ser vinculados a una coordenada geográfica, quedando excluidos del análisis espacial inicial. Esta considerable brecha de información, sintetizada en la Tabla 2, justifica la implementación de modelos LLM con flexibilidad semántica para normalizar y recuperar datos vitales para la vigilancia epidemiológica.

Tabla 2: Resumen del diagnóstico de direcciones.

Estado	Registros (%)	NP
Geocodificación exitosa	5,167 (33.02 %)	Procesadas por motor determinista
Fallo de geocodificación	10,478 (66.97 %)	Requiere normalización con LLM
Total analizado	15,645 (100 %)	Muestra de datos RPCMP-Pasto

Fuente: Elaboración propia

Rendimiento de la Estrategia Híbrida y Validación del Sistema

Con el propósito de determinar con rigor científico el impacto de la solución planteada, el análisis comparativo se centró en contrastar el rendimiento del enfoque convencional frente al flujo de trabajo (pipeline) híbrido diseñado. Para efectos de claridad metodológica, el término “Geocoder-Base” identifica el comportamiento aislado del motor determinístico original. En contraposición, las etiquetas de los modelos evaluados (por ejemplo, Llama-3.2-3B-Instruct o Gemma-3-270m-it) representan la ejecución integral del pipeline, el cual unifica el preprocesamiento estructurado, la búsqueda determinística inicial y, de forma exclusiva ante eventos de fallo, la derivación automatizada al modelo de lenguaje optimizado mediante la técnica LoRA [24].

Configuración del Dataset de Evaluación

A fin de asegurar la validez del experimento, a partir del universo de 15,645 registros del RPCMP, se consolidó un conjunto de 15,492 direcciones únicas idóneas para los procesos de aprendizaje y pruebas. Estas instancias fueron normalizadas y categorizadas manualmente por expertos, segmentando el dataset global bajo un esquema estándar de aprendizaje supervisado:70 % para entrenamiento (10,844 direcciones), 15 % para validación (2,324 direcciones) y 15 % para prueba (2,324 direcciones).

Es importante subrayar que la evaluación analítica expuesta en esta sección se ejecutó de manera estricta sobre el conjunto de prueba (n = 2,324), el cual agrupa los registros con los niveles más críticos de ambigüedad léxica, permitiendo confrontar directamente las predicciones del sistema frente a la verdad fundamental (ground truth) estandarizada.

Con el propósito de descartar cualquier tipo de sesgo y garantizar una evaluación objetiva del desempeño, el conjunto de prueba (test set) se mantuvo completamente aislado e independiente durante todo el ciclo de desarrollo[29]. Este conjunto no intervino en la fase de construcción ni en la depuración del conjunto de referencia (ground truth), y sus instancias no fueron expuestas de ninguna manera durante el proceso de ajuste fino (fine-tuning) mediante LoRA ni en la selección de hiperparámetros basada en validación. De este modo, la evaluación final reflejó de manera rigurosa la capacidad de generalización de los modelos ante datos completamente inéditos.

Eficacia en la Clasificación Estructura

El primer vector de análisis se orientó a cuantificar la precisión del sistema para clasificar de manera correcta la tipología estructural de cada dirección, distinguiendo entre las categorías de “Malla Vial”, “Manzana y Casa” o “Desconocido/ Incompleto”. En la Tabla 3 se desglosan de forma detallada las métricas estadísticas estándar de clasificación (tales como precisión, recall y puntuación F1) obtenidas para cada una de las configuraciones y arquitecturas evaluadas.

Tabla 3: Comparativa de métricas de rendimiento sobre el dataset de prueba.

Modelo	Exactitud	Precisión	Exhaustividad	F1-Score
Geocoder-Base	94.6%	93.4%	95.8%	94.2%
Gemma 3-270m-it	99.7	99.6%	99.6%	99.6%
Gemma 3-1b-it	99.2%	99.0%	99.0%	99.0%
Gemma 3-4b-it	99.5%	99.4%	99.5%	99.4%
Llama-3.2-1B-Instruct	99.6%	99.6%	99.5%	99.6%
Llama-3.2-3B-Instruc	99.6%	99.5%	99.5%	99.5%

Fuente: Elaboración propia

A partir de los datos consolidados en la Tabla 3, se evidencia que la implementación del flujo de trabajo híbrido supera de manera sistemática el rendimiento de la línea base. Resalta especialmente la integración del modelo Gemma-3-270m-it, cuya incorporación logra elevar la puntuación F1 desde 0.942 (registrado por el Geocoder-Base) hasta un óptimo 0.996. Este comportamiento resulta particularmente significativo si se considera la eficiencia computacional de dicha arquitectura, la cual

alcanza niveles críticos de precisión a pesar de contar con un tamaño de parámetros reducido.

Con el propósito de determinar la contribución específica de cada modelo evaluado sobre el proceso cartográfico, se analizó la relación directa entre el rendimiento de normalización textual y la tasa de acierto en la geocodificación espacial final medida mediante la coincidencia con las geometrías de referencia en PostGIS.

Los hallazgos indican que la precisión en la recuperación geográfica está fuertemente acoplada a la exactitud en la resolución de entidades toponímicas locales. En particular, el modelo Gemma-3-270m-it aportó la mayor recuperación geográfica efectiva dentro del pipeline híbrido, logrando una tasa de asignación espacial correcta del 99.6% sobre el conjunto de prueba. Este comportamiento se explica porque su menor grado de sobregeneralización semántica evita alteraciones indeseadas en la sintaxis de las direcciones y puntos de interés específicos de la región de estudio. Por otro lado, las arquitecturas de mayor tamaño (Gemma-3-4b-it y Llama-3.2-3B-Instruct), si bien demostraron un rendimiento fluido en tareas lingüísticas generales, presentaron una ligera degradación en la geocodificación final debido a alucinaciones ocasionales o normalizaciones excesivas sobre abreviaturas catastrales y viales locales, requiriendo un filtrado heurístico más restrictivo en la estrategia híbrida.

En sintonía con lo anterior, el rendimiento superior obtenido por el modelo de menor escala frente a arquitecturas considerablemente más grandes puede explicarse a partir de tres hipótesis técnicas principales en el contexto del ajuste fino especializado:

1. Efectos de sobreajuste benigno (benign overfitting) y optimización local: Los modelos con menor cantidad de parámetros poseen un espacio de búsqueda más restringido, lo que, bajo una estrategia de adaptación por bajo rango (LoRA) [24] con un corpus de dominio específico acotado, facilita una convergencia rápida y un acople altamente eficiente a la nomenclatura y toponimia local. En contraste, los modelos de mayor escala poseen una capacidad de representación sobredimensionada para el volumen de datos de entrenamiento provisto, lo que puede inducir a una optimización subóptima o a una interferencia con los pesos preentrenados genéricos[30].
2. Dinámica de adaptación en LoRA y capacidad de los pesos base: En arquitecturas ligeras, la actualización de las matrices de bajo rango altera proporcionalmente una mayor proporción efectiva del comportamiento del modelo en relación con su espacio latente original. Por el contrario, los modelos de 3B y 4B parámetros poseen estructuras de atención mucho más profundas, las cuales pueden requerir un mayor volumen de datos de refuerzo o un esquema de hiperparámetros diferente para que la adaptación logre redefinir las rutas de atención sin entrar en conflicto con el conocimiento base[24].
3. Sensibilidad a la cuantización y esquemas de generación: La alineación entre la complejidad estructural del modelo y la naturaleza de las consultas espaciales del ground truth sugiere que, para tareas de extracción y clasificación de baja complejidad semántica pero alta especificidad geográfica, los modelos compactos maximizan la eficiencia de la inferencia sin sufrir de los efectos de degradación que a veces se manifiestan al aplicar fine-tuning en arquitecturas sobredimensionadas con datos escasos[31].

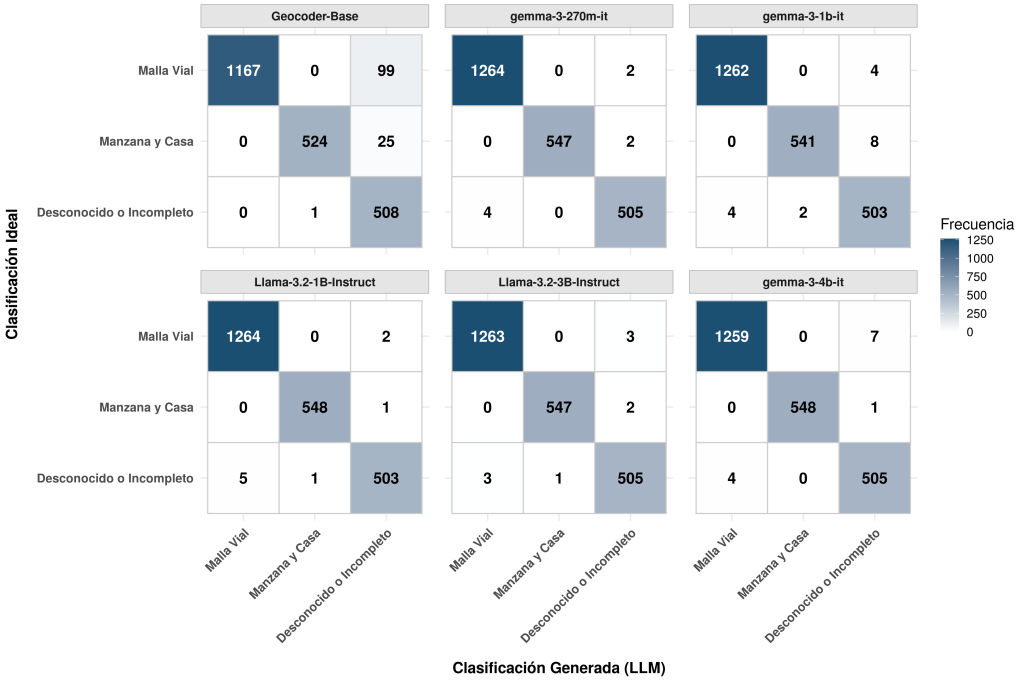

Fig. 1. *Matriz de Confusión y Métricas de Evaluación del Modelo.*
Fuente: Elaboración propia

Con el propósito de profundizar en la naturaleza de las inconsistencias corregidas por el pipeline diseñado, en la Figura 1 se presentan las matrices de confusión derivadas del proceso de clasificación, permitiendo contrastar visualmente la distribución de los aciertos y las falsas alarmas en cada categoría.

El examen detallado de la Figura 1 evidencia una limitación estructural severa en el comportamiento del Geocoder-Base, caracterizada por una marcada tendencia a la subestimación o descarte de registros potencialmente válidos. De manera puntual, el algoritmo tradicional clasificó de forma errónea un total de 99 direcciones pertenecientes a la tipología de "Malla Vial" y 25 correspondientes a "Manzana y Casa", asignándolas incorrectamente a la categoría de "Desconocido o Incompleto".

Por el contrario, la implementación del pipeline híbrido logró mitigar este sesgo de clasificación casi en su totalidad. Las arquitecturas optimizadas, tales como Llama-3.2-3B-Instruct y Gemma-3-270m-it, disminuyeron drásticamente estos falsos rechazos, restringiéndolos a un margen marginal de entre 2 y 3 registros. Este incremento sustancial en la precisión analítica permitió recuperar información espacial crítica que, bajo la lógica determinística convencional, se habría considerado como pérdida de datos nulos.

Validación de la Normalización Textual (Extremo a Extremo)

Aparte de la clasificación categórica, la efectividad real del sistema híbrido está condicionada por su competencia para reestructurar de manera exacta la cadena de texto original conforme al estándar oficial del DANE[3], paso que constituye el requisito indispensable para la georreferenciación definitiva. Con el fin de evaluar este comportamiento, en la Figura 2 se expone de manera comparativa la tasa de aciertos lograda en la coincidencia textual directa a lo largo de todo el flujo de trabajo computacional.

Los hallazgos consolidados en la Figura 2 proporcionan un sustento empírico sólido respecto a la pertinencia y robustez de la arquitectura propuesta. Mientras que el motor tradicional aislado exhibió una limitación operativa al alcanzar un tope máximo de estandarización del 84.08 %, el flujo de trabajo híbrido integrado en YACHAY-SIG logró estructurar correctamente la información en la gran mayoría de los registros de alta complejidad. Bajo este criterio de evaluación de exactitud textual estricta, el pipeline acoplado con la arquitectura Llama-3.2-3B-Instruct registró el desempeño más elevado, alcanzando un 99.40 % de efectividad en la coincidencia directa.

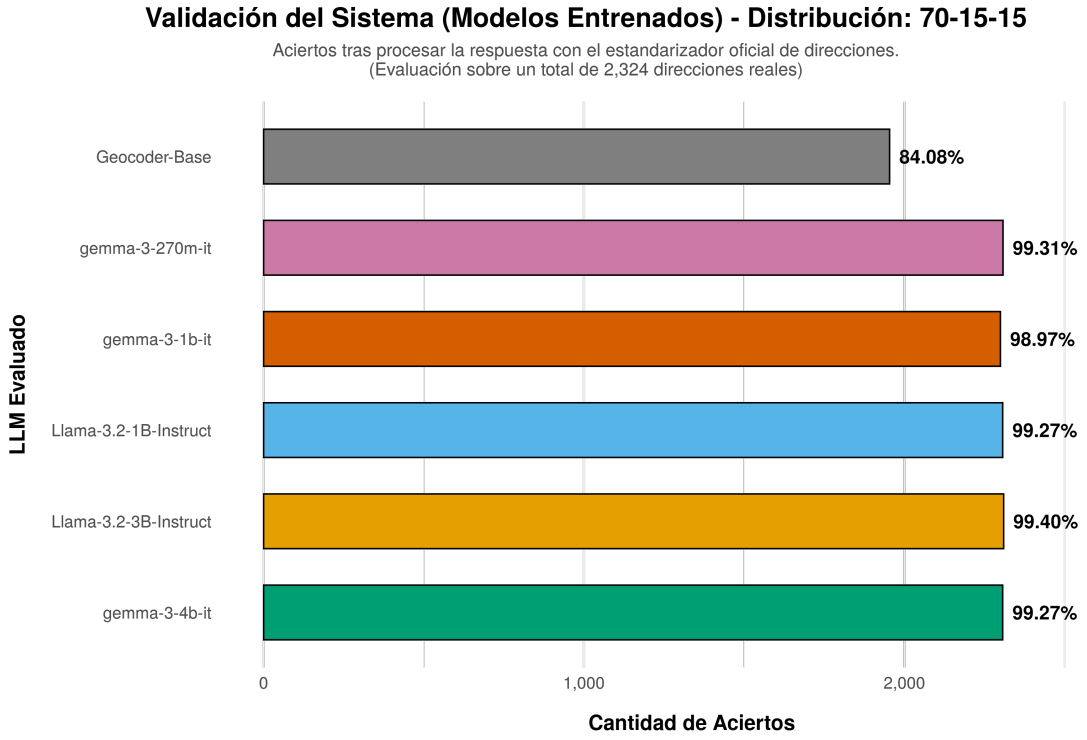

Fig. 2. *Tasa de exactitud en la normalización textual de direcciones.*
Fuente: Elaboración propia

Esta evolución en los indicadores confirma que la estrategia de procesamiento en cascada cumple un doble propósito de alta relevancia técnica: en primer lugar, actúa como un mecanismo de salvaguarda para la infraestructura computacional, al restringir la activación del modelo de lenguaje de gran escala (LLM) exclusivamente a los escenarios de fallo del Geocoder-

Base; en segundo lugar, resuelve de manera definitiva el cuello de botella tecnológico provocado por la extrema variabilidad léxica y el uso de descripciones informales en el registro oncológico, garantizando la integridad de los datos para su posterior análisis espacial.

Análisis cualitativo: Casos de éxito y errores persistentes

Con el propósito de profundizar en las fortalezas y limitaciones operativas del sistema híbrido propuesto, se llevó a cabo un análisis cualitativo de las inferencias generadas sobre el conjunto de prueba. La Tabla 4 ilustra ejemplos representativos de normalizaciones corregidas exitosamente frente a escenarios donde persistieron discrepancias o fallas en la asignación espacial final.

Tabla 4: Ejemplos representativos de normalizaciones exitosas y errores persistentes identificados en el sistema híbrido de geocodificación

Tipo de Caso	Texto de Entrada (Consulta Original)	Resultado del Sistema (Pipeline Híbrido)
Caso de Éxito 1 (Abreviatura local / Toponimia)	Cra 24 con Cl 18 sur, sector de Pandiaco	Carrera 24 # Calle 18 Sur, San Juan de Pasto
Caso de Éxito 2 (Corrección de error tipográfico)	Av los Estudiantes frente a la udenar	Avenida de los Estudiantes, Universidad de Nariño
Error Persistente 1 (Ambigüedad por homonimia)	Calle Principal, vereda Las Cochas.	Asignación imprecisa de segmento vial rural
Error Persistente 2 (Fragmentación severa)	Diagonal 5 este # 12-bis (atrás de la tienda)	Falla en la normalización de la sub-dirección

Fuente: Elaboración propia

Como se observa en la Tabla 4, en los casos de éxito, la estrategia de ajuste fino mediante LoRA permite que el modelo resuelva ambigüedades idiomáticas complejas, abreviaturas catastrales no estándar y variaciones toponímicas coloquiales propias de la región de estudio, las cuales fallaban sistemáticamente en la línea base (Geocoder-Base). No obstante, los errores persistentes se concentran primordialmente en casos límite (edge cases) caracterizados por una alta escasez de contexto espacial adyacente, direcciones con nomenclatura gravemente fragmentada o la presencia de múltiples homónimos viales en zonas rurales periféricas, donde el modelo tiende a sobregeneralizar o a requerir restricciones topológicas adicionales en PostGIS.

Impacto Geoespacial, Integración en YACHAY-SIG y Limitaciones Cartográficas

Una vez corroborada la efectividad del flujo de trabajo híbrido en el ámbito de la normalización textual, el análisis se orientó a evaluar su impacto cartográfico directo sobre la base de datos del RPCMP. Para esta fase de despliegue operativo y validación en un entorno real, el sistema integrado procesó la muestra global de 15,645 registros analizados en el diagnóstico de calidad inicial, permitiendo un contraste metodológico directo entre el desempeño del modelo propuesto y la línea base determinística.

Es importante señalar que, a diferencia de las métricas puras de Procesamiento de Lenguaje Natural (NLP), el indicador definitivo de éxito para el sistema de información geográfica se definió como la Tasa de Recuperación Geoespacial (Geocoding Match Rate). Esta métrica cuantifica de manera estricta la proporción de registros clínicos que lograron intersecularse de forma exitosa con la base de datos espacial, obteniendo así una coordenada geográfica válida bajo el sistema de referencia WGS84.

Tabla 5: Incremento de la tasa de geocodificación geoespacial tras la integración híbrida.

Fase de procesamiento	Tasa	Mejora
Línea base (solo tradicional)	33.02%	-
Estrategia híbrida (tradicional + LLM)	55.13%	+66.9%

Fuente: Elaboración propia

Tal como se desglosa en la Tabla 5, la puesta en marcha del sistema híbrido en cascada produjo un incremento significativo en la visibilidad espacial de la data epidemiológica. El pipeline propuesto logró georreferenciar de manera exitosa un total de 8,625 registros, en contraste con las 5,167 localizaciones obtenidas mediante la línea base tradicional.

Este incremento representa una transición operativa desde una cobertura inicial del 33.02 % hasta un 55.13 % de efectividad, lo que equivale a una mejora relativa del 66.9 % en la capacidad de recuperación de casos oncológicos dentro del territorio. Estos indicadores confirman la viabilidad de la arquitectura para mitigar el sesgo por pérdida de datos informales en los sistemas de registro de salud pública.

Brecha Semántica frente a Brecha Espacial

Los hallazgos expuestos revelan un fenómeno crítico de naturaleza metodológica: si bien se demostró de forma empírica que el pipeline híbrido posee la capacidad de normalizar textualmente el 99.40 % de las direcciones complejas, la tasa de éxito cartográfico final se estabilizó en el 55.13 %. Esta divergencia matemática corrobora que el algoritmo resolvió exitosamente la denominada barrera lingüística; por consiguiente, los fallos de geocodificación remanentes ya no son atribuibles a deficiencias del software o del modelo de lenguaje.

En su lugar, este comportamiento responde a una limitación estricta de la infraestructura de datos espaciales y la cartografía base del municipio. Las direcciones, a pesar de haber sido correctamente estructuradas por la IA conforme a la normativa, corresponden a zonas de expansión urbana informal, asentamientos periféricos o sectores rurales que aún carecen de

representación geométrica e indexación vial en las bases de datos cartográficas oficiales.

Validación Espacial y Visualización

Con el propósito de garantizar la integridad y la fiabilidad territorial de los datos recuperados, se implementó un protocolo de validación geométrica mediante pruebas topológicas de "punto en polígono" sobre las capas vectoriales (shapefiles) oficiales del DANE[3]. Este procedimiento permitió constatar que el 100 % de los registros médicos recuperados se localizan de manera estricta dentro de los límites político-administrativos del municipio de Pasto, eliminando la presencia de "puntos huérfanos" o coordenadas erróneas fuera del área de estudio.

La integración definitiva de estas funcionalidades en el módulo YACHAY-SIG dota al sistema de la capacidad de identificar clústeres de incidencia oncológica y patrones de distribución espacial en sectores de alta vulnerabilidad que antes carecían de visibilidad técnica. En las Figuras 3 y 4 se ilustra esta distribución a través de mapas de densidad de puntos y mapas de calor (kernel density), consolidando a la plataforma como un insumo tecnológico y estratégico de alto valor para la toma de decisiones informadas en salud pública y epidemiología[4].

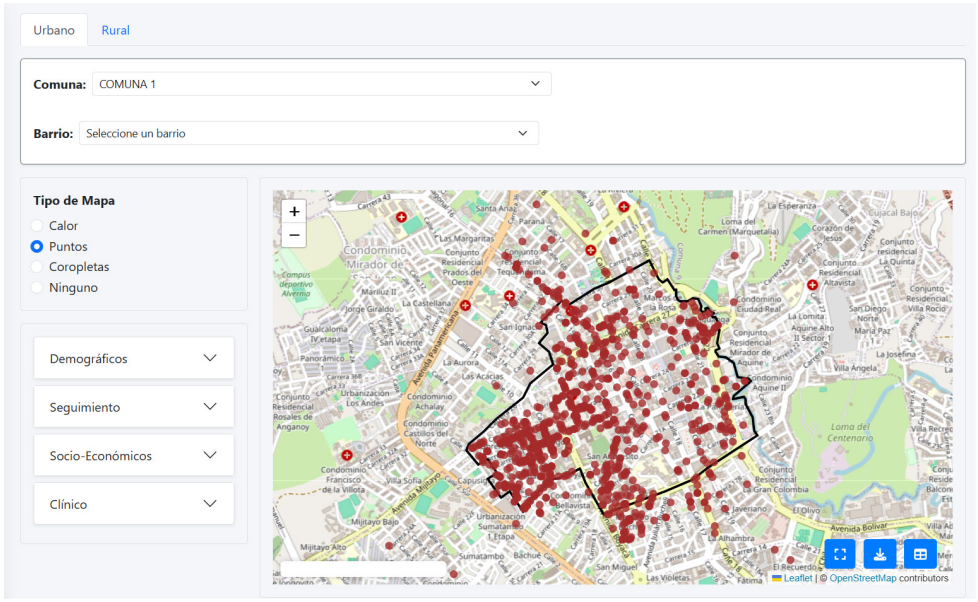

Fig. 3. *Mapa de puntos de la distribución de casos de cáncer en Pasto.*
Fuente: Elaboración propia

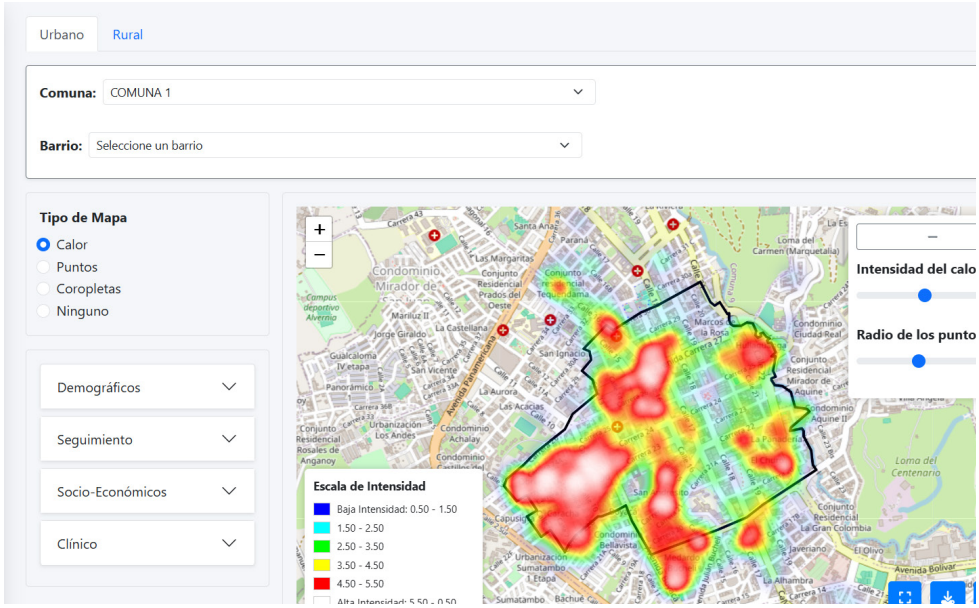

Fig. 4. *Mapa de calor de la distribución de casos de cáncer en Pasto.*
Fuente: Elaboración propia

Discusión de Resultados y Limitaciones

La precisión alcanzada en este estudio posiciona a la plataforma YACHAY-SIG como una alternativa competitiva en la geocodificación inteligente para entornos con limitaciones de datos. Al contrastar el procesamiento de lenguaje natural con la validación cartográfica, se evidencia que la arquitectura híbrida optimiza el rendimiento computacional y contribuye a mitigar el impacto del sesgo de información geoespacial en el registro oncológico evaluado.

En concordancia con lo expuesto por Guermazi et al. [13] en el ámbito logístico, los resultados corroboran que las arquitecturas neuronales superan a los algoritmos de string matching tradicionales ante datos ruidosos. Mientras que el geocodificador determinístico base alcanzó un techo operativo del 33.02 %, la integración del pipeline con el agente LLM (Llama-3.2-3B-Instruct) elevó la exactitud de estandarización semántica al 99.40 %. Esto permitió escalar la tasa global de geocodificación al 55.13 %, representando una mejora relativa del 66.9 % en la recuperación de datos (Tabla 5). Estas cifras respaldan la utilidad de transitar de sistemas rígidos hacia modelos con flexibilidad semántica en este dominio.

Asimismo, la implementación de técnicas paramétricamente eficientes como LoRA [24] demuestra la viabilidad técnica de adaptar modelos de lenguaje de propósito general a dominios especializados bajo restricciones de recursos computacionales. En el contexto de Pasto, este ajuste fino facilitó el procesamiento de nomenclaturas de tipo “barrio, manzana y lote”, un desafío técnico donde los métodos convencionales suelen presentar limitaciones para interpretar la variabilidad léxica y el contexto urbano de la región.

Desde la perspectiva de la salud pública, autores como McDonald et al. [11] sostienen que la calidad del dato geográfico en oncología depende frecuentemente de correcciones manuales para mitigar sesgos espaciales. En este sentido, la estrategia híbrida de YACHAY-SIG muestra la capacidad de automatizar la interpretación de direcciones con una tasa de error residual en la normalización de 0.60 %. Al procesar el conjunto de registros analizado, el sistema contribuye a reducir la omisión territorial de sectores que presentan dificultades en los mapas epidemiológicos tradicionales.

Un hallazgo analítico de valor para la planeación urbana es que la limitación de la tasa final de geocodificación (55.13 %) no radica en la capacidad de comprensión textual del modelo (99.40 %), sino en la ausencia de geometrías y vías específicas en la cartografía oficial de referencia. Esto evidencia una restricción en la infraestructura de datos espaciales (IDE) disponible, marcando un punto de atención para futuras actualizaciones catastrales.

A pesar del rendimiento favorable obtenido por el sistema híbrido, es preciso señalar algunas limitaciones metodológicas y técnicas que abren camino para futuras líneas de investigación. En primer lugar, los resultados evidencian una fuerte dependencia de la calidad y completitud de la infraestructura de datos espaciales (IDE) oficial y de la cartografía municipal disponible. Como se observó en el análisis de geocodificación, la falta de geometrías actualizadas o de tramos viales específicos en fuentes catastrales limita el potencial de recuperación espacial final, independientemente de la alta precisión métrica alcanzada en la estandarización semántica del texto.

En segundo lugar, el alcance de la validación experimental se circunscribió al dominio y contexto territorial específico de la región de estudio, sin incorporar pruebas sobre conjuntos de datos externos o de otras municipalidades con estructuras toponímicas heterogéneas. Si bien la estrategia de ajuste fino mediante LoRA demostró robustez local, la generalización de la arquitectura hacia otras regiones requerirá la validación con corpus heterogéneos adicionales para descartar posibles sesgos de adaptación a las particularidades léxicas del área evaluada.

Finalmente, la integración de estos componentes mediante herramientas de código abierto como PostgreSQL y Metabase ilustra un esquema operativo funcional para el proyecto YACHAY. Este modelo de arquitectura en cascada ofrece un enfoque metodológico útil para ser analizado y adaptado por otros registros poblacionales que enfrenten desafíos similares en la estandarización de su nomenclatura domiciliaria.

CONCLUSIONES Y TRABAJOS FUTUROS

La presente investigación muestra que la integración de Modelos de Lenguaje de Gran Escala especializados mediante la técnica de Adaptación de Bajo Rango (LoRA) constituye una alternativa eficaz para la normalización de datos geográficos heterogéneos en el dominio evaluado. Al superar las limitaciones de los métodos determinísticos convencionales los cuales registraron una cobertura base de apenas el 33.02 %, el pipeline híbrido respaldado por la arquitectura ajustada alcanzó una exactitud textual del 99.40 % en la interpretación de nomenclaturas complejas, recuperando información epidemiológica que previamente requería intervención manual.

La implementación de una estrategia de procesamiento híbrida en cascada demostró ser un enfoque eficiente para el manejo de los registros. Al emplear el geocodificador tradicional como primera línea de procesamiento y activar el modelo inteligente ante registros ambiguos, se optimizó el uso del recurso computacional. Esta sinergia permitió elevar la Tasa de Recuperación Geoespacial al 55.13 % (un incremento relativo del 66.9 %), indicando que el modelo contribuyó a resolver la barrera

semántica. Este hallazgo evidencia que un factor limitante relevante para lograr una georreferenciación completa radica en la cobertura y actualización de la cartografía oficial disponible en zonas de expansión informal y áreas rurales del municipio de Pasto.

Finalmente, este trabajo ratifica la utilidad metodológica del uso de modelos de pesos abiertos y el despliegue sobre herramientas de código abierto como PostgreSQL y PostGIS, ofreciendo una arquitectura replicable para ser analizada por otros registros poblacionales o sistemas de vigilancia epidemiológica que afronten desafíos similares en la estandarización de su nomenclatura domiciliaria.

Como trabajos futuros, el horizonte de esta investigación se encamina hacia la implementación operativa del sistema desarrollado. Como paso inmediato, se contempla el despliegue de la plataforma YACHAY-SIG en el entorno del Registro Poblacional de Cáncer del Municipio de Pasto (RPCMP). Con su operación, se espera que los procesos de normalización y georreferenciación automática descritos contribuyan a la generación de cartografía epidemiológica orientada a apoyar la identificación de patrones espaciales y a optimizar los tiempos de gestión institucional.

Asimismo, la investigación proyecta la incorporación de técnicas de geocodificación inversa dentro del pipeline de procesamiento para evaluar la coherencia topológica entre las coordenadas obtenidas y las descripciones textuales originales, sirviendo como un mecanismo complementario de control de calidad y auditoría espacial.

Finalmente, se planea explorar la transferencia tecnológica y la adaptación de los modelos en otros contextos regionales de Colombia, con el propósito de evaluar su desempeño ante nuevas estructuras toponímicas y aportar al fortalecimiento de los sistemas de información espacial en salud pública.

REFERENCIAS

[1] R. Timarán, L. Bravo, A. Hidalgo, R. Suárez, A. Chaves, and F. Vidal, “YACHAY-SIG: Sistema georreferenciado para el registro poblacional de cáncer del municipio de Pasto”, in Avances en Investigaciones de Sistemas Inteligentes y Gestión de Conocimiento: Libro de Resúmenes del CISIGESCO 2025. Popayán, Colombia: Editorial Institución Universitaria Colegio Mayor del Cauca, 2025, p. 11. [Online]. Available: https://sired.udenar.edu.co/17622/1/17622.pdf

[2] R. Timarán, A. Torres, F. Vidal, and A. Hidalgo. “YACHAY: un sistema inteligente para la gstión de la incidencia de cáncer en el municipio de Pasto”, in Proc. Encuentro Internacional de Educación en Ingeniería ACOFI (EIEI 2025), Cartagena, Colombia, Sep. 16-19, 2025.

[3] Departamento Administrativo Nacional de Estadística (DANE), “Geoportal DANE: Marco Geoestadístico Nacional,” 2026. [Online]. Available: https://geoportal.dane.gov.co/

[4] J. D. Viveros, R. Timarán and A. Calderón, “Diagnóstico de la Calidad de Datos para la Geocodificación Inteligente en el Registro Poblacional de Cáncer de Pasto”, in Avances en Investigaciones de Sistemas Inteligentes y Gestión de Conocimiento: Libro de resúmenes del CISIGESCO 2025, Popayán, Colombia: Editorial Institución Universitaria Colegio Mayor del Cauca, 2025, p. 19. [Online]. Available: https://sired.udenar.edu.co/17622/1/17622.pdf

[5] R. Timarán, G. Hernández and N. Quemá, “Geocodificador de eventos delictivos georreferenciados a nivel de direcciones urbanas en el municipio de Pasto”, en Memorias del 5° Congreso Internacional de Gestión Tecnológica y de la Innovación (COGESTEC), Bucaramanga, Colombia: Universidad Industrial de Santander, 2016, pp. 314-327.

[6] J. Hu, et al., “GeoAI agent: To empower LLMs using geospatial tools for address standardization and spatial analysis,” *International Journal of Geographical Information Science*, vol. 39, no. 4, pp. 612–635, 2025.

[7] M. Goodchild, “GIScience and systems: Past, present, and future trajectories in geographic automation,” Computers, Environment and Urban Systems, vol. 102, p. 101954, 2023.

[8] A. Zhai, et al., “Transformer-based models for spatial entity extraction and address matching: A comparative review,” *ISPRS International Journal of Geo-Information*, vol. 12, no. 7, p. 284, 2023.

[9] K. Yao, et al., “Contextual semantic understanding in street address parsing using pre-trained language models,” *Transactions in GIS*, vol. 27, no. 5, pp. 1420–1440, 2023.

[10] R. Mai, et al., “Mapping with words: Integrating large language models into geospatial practice and location intelligence,” ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XI-4/W2-2026, pp. 115–122, 2026.

[11] Y. J. McDonald, M. Schwind, D. W. Goldberg, A. Lampley y C. M. Wheeler, “An analysis of the process and results of manual geocode correction,” *Geospatial Health*, vol. 12, no. 1, 2017, doi: 10.4081/gh.2017.526.

[12] S. Gupta y K. Nishu, “Mapping Local News Coverage: Precise location extraction in textual news content using fine-tuned BERT based language model,” en Proc. of the Fourth Workshop on Natural Language Processing and Computational Social Science, 2020, pp. 155-162.

[13] Y. Guermazi, S. Sellami y O. Boucelma, “A RoBERTa Based Approach for Address Validation,” en New Trends in Database and Information Systems, Springer, 2022, doi: 10.1007/978-3-031-15743-1_31.

[14] W. X. Zhao et al., “A Survey of Large Language Models”, *CoRR*, vol. abs/2303.18223, 2023 [Online]. Available: https://arxiv.org/abs/2303.18223

[15] Meta AI, “Llama 3.2: Model Cards and Prompt formats,” 2025. [En línea]. Disponible en: https://ai.meta.com/blog/llama-3-2-

connect-2024/

[16] M. Batty, "The computational city: Big data and geographical information systems in urban modeling," *Progress in Human Geography*, vol. 47, no. 2, pp. 210–228, 2024.

[17] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Geographic Information Science and Systems, 4th ed. Hoboken, NJ: Wiley, 2023.

[18] T. Vaswani et al., "Attention is all you need," *in Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[19] J. L. Pérez et al., "A review of hybrid geocoding architectures for health informatics: Challenges in scalability and data governance," *International Journal of Health Geographics*, vol. 23, no. 1, p. 14, 2024.

[20] M. Open-Weights Consortium, "Efficient parameter fine-tuning of open-source language models for geospatial parsing in resource-constrained environments," Computers, Environment and Urban Systems, vol. 118, p. 102271, 2025.

[21] Gemma Team et al., "Gemma 3 Technical Report," arXiv preprint arXiv:2503.19786, 2025, doi: 10.48550/arXiv.2503.19786.

[22] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," *in Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 8024–8035, 2019.

[23] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo, "LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models," *in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vol. 3: System Demonstrations, Bangkok, Thailand, Aug. 2024, pp. 400-410, doi: 10.18653/v1/2024.acl-demos.38.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[25] T. Wolf et al., "Hugging Face's transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[26] G. Gerganov, "llama.cpp: Port of Facebook's LLaMA model in C/C++," GitHub Repository, 2023. [En línea]. Available: https://github.com/ggerganov/llama.cpp

[27] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," International Conference on Learning Representations (ICLR), 2017.

[28] L. Prechelt, "Early stopping—but when?", in Neural Networks: Tricks of the Trade, Berlin, Heidelberg: Springer, 2012, pp. 53–67.

[29] C. M. Bishop, Pattern Recognition and Machine Learning, New York: Springer, 2006.

[30] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 2022.

[31] Y. Wang, Z. Yu, W. Yao, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang,2R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, and Y. Zhang,3"PandaLM: An Automatic Evaluation Benchmark for LLM Instruction4Tuning Optimization," in International Conference on Learning Representations (ICLR), 2024.

FINANCIACIÓN

Este proyecto es financiado con recursos del Sistema de Investigaciones de la Universidad de Nariño.

CONFLICTOS DE INTERÉS

No existen

CONTRIBUCIÓN DE AUTORÍA

Redacción - borrador original: Silvio Ricardo Timarán Pereira, Jonathan David Viveros Córdoba.

Redacción - revisión y edición: Silvio Ricardo Timarán Pereira, Jonathan David Viveros Córdoba.