ISSN: 1692-7257 - Volume 2 – Issue 48 - 2026

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

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

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

Silvio Ricardo Timarán Pereira[1]

Jonathan David Viveros Córdoba[1]

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

Received: 09 April 2026. Revised: 22 June 2026. Accepted: 24 July 2026. Published: 28 July 2026.
Corresponding Author: Silvio Ricardo Timarán Pereira

How to cite: 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

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract:

The accuracy of geospatial analysis within the Population-based Cancer Registry of the Municipality of Pasto (RPCMP) is severely compromised by the heterogeneous and ambiguous nature of physical addresses collected manually. These source data inconsistencies limit the analytical reach of the YACHAY-SIG infrastructure for generating multiscale territorial intelligence. This paper describes the development and implementation of a hybrid methodological framework designed to optimize spatial data integrity through a two-phase processing workflow.The initial phase deploys a deterministic geocoding engine specifically tailored to the local urban morphology. For records exhibiting high semantic complexity such as non-conventional "block and lot" (manzana y lote) nomenclatures a second stage based on Large Language Models (LLMs) is incorporated. This component integrates Natural Language Processing (NLP) techniques, utilizing TF-IDF and Word2Vec vectorization for the discernment and correction of structural anomalies in the addresses.To maximize efficiency in resource-constrained computational environments, a Supervised Fine-Tuning (SFT) protocol was applied using the Low-Rank Adaptation (LoRA) technique on several open-source models. Experimental results reveal that the integration of the Llama-3.2-3B-Instruct model achieved superior performance, reaching a normalization accuracy of 99.40%. System evaluation over an effective test dataset of 2,324 addresses (extracted from a total corpus of 15,492 unique instances allocated into 70% for training, 15% for validation, and 15% for testing) using standard metrics including precision, recall, and F1-score demonstrates a substantial improvement in spatial recovery, increasing from an initial baseline of 33.02% to 55.13%, effectively mitigating the semantic gap regardless of local cartographic constraints. This architecture stands as a scalable data engineering model for epidemiological registries, with future work focusing on integration with spatial providers such as OpenStreetMap and the Agustín Codazzi Geographic Institute (IGAC) to strengthen national geographic coverage.

Keywords: data normalization, geocoding, large language models, natural language processing, supervised learning.

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.

INTRODUCTION

Globally, public oncology requires epidemiological surveillance systems with analytical capabilities that go beyond incidence recording, with priority on understanding the spatial distribution of the disease. In this setting, the Population-based Cancer Registry of the Municipality of Pasto (RPCMP) is the central node for analyzing the cancer burden in the region. However, turning these data into strategic assets for clinical and administrative decision-making requires high-precision georeferencing mechanisms that enable reliable multiscale visualization.

In response to this technical need, YACHAY-SIG[1]has been implemented, an intelligent geographic information infrastructure designed to strengthen the analytical capacity of the RPCMP. The platform rests on a relational data architecture managed in PostgreSQL with the PostGIS geospatial engine, which ensures efficient storage and processing of clinical and geographic variables. The system allows prevalence and incidence to be disaggregated into micro-local territorial units (neighborhoods, communes, corregimientos and veredas), supporting cartographic synthesis through heat maps and dynamic choropleth maps[2].

Nevertheless, the operational integrity of YACHAY-SIG is intrinsically tied to the quality of the input data. Patient location is derived from postal addresses captured in healthcare settings, which show high entropy and structural variability. Previous diagnoses confirm that ambiguity in nomenclature hinders coupling with the National Geostatistical Framework of the National Administrative Department of Statistics (DANE) [3], the official cartographic reference base [4]. The problem is worse in the city of Pasto, where official nomenclatures coexist with informal descriptors based on “blocks and lots”, producing a significant semantic interoperability gap [5].

In recent years, the evolution of geospatial artificial intelligence (GeoAI) has transformed how unstructured spatial data are processed and normalized[6]. Traditionally, geographic information systems relied on rigid rules and heuristic methods, which showed severe limitations when faced with imprecise, colloquial or incomplete location descriptions[7]. To overcome the limitations of character-based and traditional schema-based approaches, recent research has incorporated natural language processing (NLP) techniques grounded in Transformer architectures[8]. In particular, models such as BERT and RoBERTa have been shown to significantly improve geographic entity extraction and address matching by capturing the contextual semantics of the text[8],[9].

More recently, the emergence of large language models (LLMs) has introduced a new paradigm in intelligent geocoding: coordinate prediction becomes an advanced generative task, and complex locality descriptions with relative spatial relations can be resolved[10]. These technologies open a fundamental pathway to closing the semantic gap in complex domains such as sociosanitary and population registries[6],[10].

Historically, geocoding in settings with non-standardized naming has been approached through a range of methodologies. At the regional level, deterministic engines have been developed based on string matching algorithms optimized for the local road network[5]. Although these systems provide a solid foundation, their performance drops exponentially when they face unstructured records, which are common in peripheral and rural areas. In parallel, the international scientific literature indicates that manual processing of these coordinates is economically unfeasible given large data volumes[11], which has driven the adoption of deep learning models and Transformer architectures (such as BERT and RoBERTa) for resolving textual ambiguities [12], [13].

At the current technological frontier, Artificial Intelligence has converged toward Large Language Models (LLMs), which show disruptive capabilities in Natural Language Processing (NLP)[14]. However, specializing foundation models for critical geographic tasks requires specific adaptation protocols. In this context, the Low-Rank Adaptation (LoRA) technique has become established as the standard for fine-tuning these models, enabling cutting-edge architectures, such as Gemma 3 and Llama 3.2, to reach higher levels of precision in address normalization[15].

Despite the methodological advances described, a significant scientific gap persists in the automated normalization of addresses with high entropy[16]. Traditional approaches, based on deterministic algorithms and heuristic string matching rules, assume rigid urban naming, and they break down operationally when faced with severe typographical errors, informal abbreviations, or the absence of essential normative components[17]. With the adoption of classic Transformer architectures, prior research partially mitigated this problem through entity extraction based on contextual semantics[8],[9]; however, these architectures are severely limited by their discriminative nature, which prevents them from solving complex generative tasks or interpreting relative spatial relationships expressed in colloquial language (for example, descriptions such as “behind the main park, vacant lot”)[18].

The LLM-based strategy addresses this gap through a paradigm shift, turning normalization into a structured generative task capable of reasoning about spatial ambiguity [10]. Nevertheless, the current literature lacks efficient hybrid frameworks that integrate the deterministic precision of traditional geographic information systems with the semantic flexibility of language models fine-tuned using computationally low-cost techniques (LoRA)[13], a fundamental gap that

the present work seeks to address.

Compared with other hybrid proposals reported in the literature, which generally amount to chaining black-box commercial geocoding services with generic text-cleaning heuristics[19], the distinguishing contribution of this research lies in the design of a tightly coupled framework optimized for regional contexts with limited resources. Rather than relying on proprietary massive models, the proposed strategy integrates a local deterministic engine in PostgreSQL/PostGIS with an open-weight language agent specialized through efficient parameter adaptation (LoRA)[20]. This approach not only guarantees the sovereignty and privacy of the sensitive epidemiological data of the Population-based Cancer Registry, but also operates with high precision on complex informal nomenclatures ("blocks and lots") characteristic of intermediate Latin American cities, achieving an optimal balance between computational cost, spatial interpretability, and coordinate retrieval rate[19],[20].

This article presents the implementation of a hybrid strategy that combines the robustness of a deterministic geocoding engine with the semantic flexibility of an NLP agent powered by optimized LLMs. This architecture strengthens the integrity of epidemiological surveillance and provides a solid scientific basis for oncological analysis with a territorial approach.

MATERIALS AND METHODS

Materials

The design and implementation of the address normalization module, and its subsequent incorporation into the YACHAY-SIG platform, was structured around three essential components: official data repositories, state-of-the-art large language models (LLM), and a dedicated high-capacity computing infrastructure for processing.

Data Sources and Cartographic References

As a primary input, the historical records of the Population-based Cancer Registry of the Municipality of Pasto (RPCMP) were used, which consolidate the clinical and sociodemographic data of the population diagnosed in the region.

To guarantee the reproducibility of the study, the initial dataset provided by the RPCMP consisted of a total of 15,645 raw geographic records of oncological patients. During the initial audit and data cleansing phase, rigorous quality control filters were applied. Specifically, 153 records were excluded: entries with null or corrupt geographic coordinates outside the urban perimeter (n = 68), exact duplicates generated by multiple care admissions of the same patient (n = 45), and postal descriptions with extreme entropy or devoid of useful information (n = 40). After this systematic cleansing process, a curated corpus of 15,492 valid records was obtained, which were partitioned in a stratified manner for the experimental training, validation, and test subsets (see Table 1).

Table 1: Partition of the RPCMP Dataset

Category / Subset	No. Rec.	Exclusion / inclusion criteria	Composition / Distribution
Initial Raw Records	15,645	Total extraction from the RPCMP database.	All raw oncology notifications
Excluded Records (Cleaning)	(153)	Null coordinates or coordinates outside the municipality (n=68). Exact duplicates (n=45). Extreme entropy / empty fields (n=40)	Records discarded due to geocoding infeasibility
Final Curated Dataset	15,492	Records suitable for processing and normalization.	100% of the clean data integrated into the hybrid pipeline
Training Subset	10,844 (70%)	Assigned for fine-tuning using LoRA.	Stratified distribution by commune and severity
Validation Subset	2,324 (15%)	Used for hyperparameter monitoring and early stopping.	Independent representative sample.
Test Subset (Test)	2,324 (15%)	Reserved for blind evaluation of final performance.	Complex, high-entropy rural cases.

Source: Own elaboration

To ensure precision in spatial localization and the geocoding process, two key cartographic instruments were articulated:

National Geostatistical Framework (MGN): Provided by DANE[3], this resource allowed us to establish the official political-administrative structure, from the division into urban neighborhoods and sectors to the rural organization into corregimientos and veredas.

Pasto Geocoding Engine: We integrated the open-source tool developed by Timarán et al.[5], which uses string matching algorithms to convert physical addresses into geographic coordinates. This engine is essential because it is calibrated specifically for the road topology and nomenclature of San Juan de Pasto.

Large Language Models (LLMs)

In order to carry out the experimentation and fine-tuning phases, five variants belonging to two families of open-weights models were selected. The choice of these models responded to the need to balance processing power with operational efficiency in hardware architectures with limited capabilities:

Gemma 3 IT series (270M, 1B and 4B): Developed by Gemma Team[21], these models represent the state of the art in

lightweight architectures. They are characterized by optimizing logical reasoning and compliance with complex instructions while keeping energy and memory consumption low.

Llama 3.2 series (1B-Instruct and 3B-Instruct): We implemented these compact versions from Meta AI[15], which have been refined for high-performance deployments. They stand out for their advanced semantic understanding and for ensuring low latency in limited computational environments.

Software Ecosystem and Development Tools

To guarantee scientific replicability and promote technological sovereignty, the software infrastructure was built entirely from open-source tools. The components of the ecosystem are detailed below:

Spatial Data Management: PostgreSQL was adopted as the relational database management system, complemented with the PostGIS extension for the administration, storage and querying of geometries referenced in the WGS84 coordinate system.

Programming Environment and Dependencies: The system logic was developed in Python[22], using the uv package manager to ensure efficient and isolated dependency management.

Training and Fine-Tuning: The training and optimization phases of the models were executed using the LLaMA Factory [23] and Unsloth tools in combination. Model adaptation was based on the efficient parameterization technique LoRA (Low-Rank Adaptation)[24], run on high-performance graphics infrastructure with CUDA support.

Results Visualization: The spatial representation of the information was integrated into the YACHAY-SIG web interface[1], a platform built on modern web architectures for the generation and visual analysis of dynamic heat maps and choropleth maps[2].

Hardware Infrastructure and Deployment

The training and fine-tuning process of the models relied on a high-performance workstation equipped with an NVIDIA A100 Tensor Core graphics processing unit (GPU) (with 40 GB of VRAM) compatible with CUDA technology. The computational workflow ran on the Ubuntu 22.04 LTS operating system, using Python 3.10.12, PyTorch 2.3.0 [22] and the Hugging Face Transformers library[25], with LLaMA-Factory as the main interface[23].

Subsequently, to ensure technical viability and efficiency in the production deployment and inference stage, the optimized models were converted and quantized to the GGUF format (Q5_K_M)[26]. This strategy allowed optimal management of computational resources and a substantial reduction in response latency, without compromising the spatial precision obtained during training.

Methods

The methodological framework of this study is based on an applied research approach, organized through a series of sequential and articulated stages. This structure allows a rigorous transition from the diagnosis and treatment of primary data sources to the functional validation of the intelligent system once it is integrated into the global architecture.

Data Diagnosis and Characterization Phase

During the initial stage, a detailed examination of the database corresponding to the Population-based Cancer Registry of the Municipality of Pasto (RPCMP) was carried out. Using Natural Language Processing (NLP) techniques, inconsistencies in the records were systematically analyzed, and critical error patterns were detected, such as the use of unconventional abbreviations, the omission of grammatical connectors and the recurrence of address systems based on block-and-lot schemes[4].

This exploratory analysis was essential for defining the specific training parameters and requirements, and it guided the fine-tuning of the language models toward resolving the lexical and structural particularities identified in the local data.

Corpus Construction for Fine-Tuning

To specialize the language model (LLM), a structured training dataset was developed through two fundamental processes:

Manual Labeling (Ground Truth): An exhaustive normalization procedure was carried out in which domain experts manually analyzed and standardized each record in accordance with the official address regulations. This process made it possible to establish the “ground truth,” which is indispensable for guaranteeing the quality of the model’s learning.

Instruction Formatting (Instruction Tuning): The data were organized under an instruction scheme designed for supervised learning. This stage consisted of creating logical Input (address with noise or inconsistencies) -> Output (normalized address and structural classification) pairs. This structure facilitated the preparation of the data for the Continuous Supervised Training (SFT) phase, optimizing the model’s ability to interpret and correct complex urban nomenclature.

Configuration of Fine-Tuning via LoRA

The central component of the methodology was the implementation of the Low-Rank Adaptation (LoRA) technique[24]. Instead of fully retraining the parameters of the Gemma 3 and Llama 3.2 models, low-rank matrices were inserted into all of their linear layers. This technical procedure offered three fundamental advantages:

Resource Optimization: It enabled a drastic reduction in VRAM memory consumption by freezing the original weights and updating only the adaptation parameters.

Preservation of Capabilities: It facilitated the model's specialization in the local semantics and nomenclature of the city of Pasto, preventing the degradation of the general reasoning capabilities inherent to the base LLM.

Computational Efficiency: Training was performed in bfloat16 mixed-precision format (bf16), optimized through the use of Gradient Checkpointing and Flash Attention.

Operationally, the training process was configured for a total of 3 epochs with an effective batch size of 16 (using a per-device batch size of 4 and gradient accumulation of 4 steps) to optimize stability in weight updating. An initial learning rate of 2×10^{-4} was implemented with a cosine learning rate scheduler and a warmup ratio of 0.03[27].

The validation strategy evaluated the model's performance at the end of each epoch using the validation subset, applying an early stopping criterion based on minimizing validation loss with a patience margin of 2 epochs to prevent overfitting[28]. For the specific LoRA parameterization, a rank factor (r) of 16, a scaling factor (α) of 32, and a dropout rate of 0.05 applied to the attention projections were established[23].

Implementation of the Hybrid Geocoding Algorithm

The transformation of textual descriptions into spatial coordinates was based on a bidirectional, sequential workflow designed to maximize the record recovery rate.

Phase 0, Textual Preprocessing and Cleaning

Before the geospatial evaluation, the original text strings pass through a structured cleaning module. Using regular expressions, the algorithm standardizes capitalization, removes anomalous punctuation, and systematically separates concatenated alphabetic and numeric characters (for example, converting "24A" into "24 A"). This preliminary stage is critical for mitigating semantic noise before any normalization or search process.

Phase 1, Enhanced Deterministic Engine (Proximity Search)

At this stage, the process employs a base geocoder[5], which was strengthened through advanced heuristics and greater syntactic flexibility. Unlike the original strict-order approach, the new algorithm can interpret inverted or partial nomenclatures. In addition, the exact-match criterion was replaced by a fuzzy matching system with a confidence scale from 0 to 1.

Under this scheme, if an address is not located literally in the road network, the system evaluates adjacent locations by applying penalty coefficients for numerical distances or variations in alphabetic suffixes. This makes it possible to assign the coordinates of the closest possible location, provided that a predefined metric confidence threshold is met.

Phase 2, Intelligent Recovery via LLM

As the last processing instance, records that present severe ambiguities, extreme colloquial references, or that result in a total failure of the deterministic engine are forwarded to the previously fine-tuned language model (LLM). In this phase, the LLM operates as a semantic translator capable of inferring the geographic intent of the text and reconstructing the address under the official DANE standards.

Once normalized, the string is reinjected into Phase 1. This second attempt allows the deterministic engine to process a predictable structure, achieving successful georeferencing in the WGS84 system for data that, under conventional methods, would be discarded because of their low quality.

To validate the performance of the hybrid system, an evaluation protocol was established based on two complementary dimensions: the linguistic precision of the normalization and the operational effectiveness of the geocoding.

Evaluation Protocol and Success Metrics

Linguistic Evaluation (Normalization Quality)

To determine the technical soundness of the fine-tuned models (Gemma 3 and Llama 3.2), the accuracy metric was used. It was calculated by directly comparing the addresses normalized by the model in the test dataset with the ground truth established through expert manual labeling. This evaluation made it possible to measure the models' ability to abstract textual noise, correct spelling errors, and restructure the information under the official nomenclature standards defined by DANE[3].

Effectiveness evaluation (recovery rate)

Because the final spatial precision is conditioned by cartographic coverage and the deterministic engine, the main metric of the hybrid system was defined as the geocoding match rate.

This indicator represents the percentage of addresses or geographic records that the system manages to locate and correctly associate with their corresponding spatial coordinates or reference points within the official geographic database. Operationally, a high value indicates that the model can process and structure imprecise or incomplete location descriptions from citizens, successfully translating them into digital maps without significant loss of information.

This metric quantifies the operational impact of the pipeline by measuring the proportion of medical records that, having initially failed in the traditional deterministic engine, obtain a valid coordinate after being processed by the hybrid flow (semantic normalization with LLM and fuzzy search). This indicator therefore evaluates not only the model's interpretive capacity but also the net increase in the system's geospatial coverage, acknowledging the interdependence between semantic improvement and the available cartographic infrastructure.

RESULTS AND DISCUSSION

The analysis of the implemented strategy made it possible to corroborate the effectiveness of large language models (LLM) as tools for the recovery of geospatial data essential for the RPCMP. The findings are detailed below, structured into the quality diagnosis, the performance of the fine-tuning process, and the impact obtained on final georeferencing.

Diagnosis of Data Quality in the RPCMP

In order to establish a baseline and gauge the technical complexity of the cancer records in the municipality of Pasto, an initial diagnosis was carried out using the open-source geocoder of Timarán et al.[5]. Although this system represents a regional technological pillar, it has a deterministic architecture designed for predictable road structures, which limits its performance with highly variable data.

The diagnosis confirmed that the inconsistencies and ambiguities present in the RPCMP constitute a critical barrier for systems based on rigid rules. Addresses captured in clinical contexts tend to present:

- Typographical errors and grammatical omissions.
- Non-standardized abbreviations.
- Colloquial descriptions unrelated to conventional string matching algorithms.

A key methodological factor was the processing of the records under the two naming models that coexist in the city: "road grid" and "neighborhood-block". An exhaustive protocol was applied in which each record was evaluated by both functions, validating the result if either of them succeeded in assigning coordinates.

The results on a sample of 15,645 records showed the scale of the challenge: only 5,167 addresses (33.02 %) were successfully geocoded. In contrast, 10,478 records (66.97 %) could not be linked to a geographic coordinate and were excluded from the initial spatial analysis. This considerable information gap, summarized in Table 2, justifies the implementation of LLM models with semantic flexibility to normalize and recover data vital for epidemiological surveillance.

Table 2: Summary of the address diagnosis.

Status	Records (%)	NP
Successful geocoding	5,167 (33.02 %)	Processed by deterministic engine
Geocoding failure	10,478 (66.97 %)	Requires normalization with LLM
Total analyzed	15,645 (100 %)	RPCMP-Pasto data sample

Source: Own elaboration

To determine with scientific rigor the impact of the proposed solution, the comparative analysis focused on contrasting the performance of the conventional approach against the hybrid pipeline designed. For methodological clarity, the term "Geocoder-Base" identifies the isolated behavior of the original deterministic engine. By contrast, the labels of the evaluated models (for example, Llama-3.2-3B-Instruct or Gemma-3-270m-it) represent the full execution of the pipeline, which unifies structured preprocessing, the initial deterministic search and, only in the event of failures, the automated referral to the language model optimized through the LoRA technique [24].

Performance of the Hybrid Strategy and System Validation

Evaluation Dataset Configuration

To ensure the validity of the experiment, a set of 15,492 unique addresses suitable for the learning and testing processes was consolidated from the universe of 15,645 RPCMP records. Experts normalized and manually categorized these instances, segmenting the global dataset under a standard supervised learning scheme: 70% for training (10,844 addresses), 15% for

validation (2,324 addresses) and 15% for testing (2,324 addresses).

The analytical evaluation presented in this section was performed strictly on the test dataset (n = 2,324), which groups the records with the most critical levels of lexical ambiguity. This allows the system's predictions to be compared directly against the standardized ground truth.

To rule out any type of bias and guarantee an objective evaluation of performance, the test dataset was kept completely isolated and independent throughout the entire development cycle[29]. This dataset did not participate in the construction phase or in the debugging of the ground truth, and its instances were not exposed in any way during the fine-tuning process using LoRA or in the validation-based hyperparameter selection. The final evaluation thus rigorously reflected the generalization capacity of the models on completely unseen data.

Effectiveness in Structure Classification

The first vector of analysis aimed to quantify the system's precision in correctly classifying the structural typology of each address, distinguishing between the categories "Road Network", "Block and House" or "Unknown/Incomplete". Table 3 details the standard statistical classification metrics (such as precision, recall and F1-score) obtained for each of the configurations and architectures evaluated.

Table 3: Comparison of performance metrics on the test dataset.

Model	Accuracy	Precision	Recall	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%

Source: Own elaboration

The data consolidated in Table 3 show that the hybrid workflow systematically outperforms the baseline. The integration of the Gemma-3-270m-it model stands out in particular: its incorporation raises the F1-score from 0.942 (recorded by Geocoder-Base) to an optimal 0.996. This behavior is especially significant given the computational efficiency of this architecture, which reaches critical levels of precision despite its reduced parameter size.

To determine the specific contribution of each evaluated model to the cartographic process, we analyzed the direct relationship between textual normalization performance and the success rate in final spatial geocoding, measured through matching with the reference geometries in PostGIS.

The findings indicate that precision in geographic retrieval is strongly coupled to accuracy in resolving local toponymic entities. In particular, the Gemma-3-270m-it model provided the greatest effective geographic retrieval within the hybrid pipeline, achieving a correct spatial assignment rate of 99.6% on the test dataset. This behavior is explained by its lower degree of semantic overgeneralization, which prevents unwanted alterations in the syntax of addresses and points of interest specific to the study region. On the other hand, the larger architectures (Gemma-3-4b-it and Llama-3.2-3B-Instruct), although they showed fluent performance in general linguistic tasks, presented a slight degradation in final geocoding due to occasional hallucinations or excessive normalizations of local cadastral and road abbreviations, requiring more restrictive heuristic filtering in the hybrid strategy.

In line with the above, the superior performance of the smaller-scale model compared with considerably larger architectures can be explained by three main technical hypotheses in the context of specialized fine-tuning:

1. Benign overfitting effects and local optimization: Models with fewer parameters have a more restricted search space, which, under a low-rank adaptation (LoRA) strategy [24] with a limited domain-specific corpus, facilitates rapid convergence and highly efficient adaptation to local nomenclature and toponymy. In contrast, larger-scale models have a representational capacity that is oversized for the volume of training data provided, which can lead to suboptimal optimization or to interference with the generic pretrained weights[30].
2. Adaptation dynamics in LoRA and capacity of the base weights: In lightweight architectures, updating the low-rank matrices proportionally alters a larger effective share of the model's behavior relative to its original latent space. By contrast, the 3B- and 4B-parameter models have much deeper attention structures, which may require more reinforcement data or a different hyperparameter scheme for the adaptation to redefine the attention pathways without conflicting with the base knowledge[24].
3. Sensitivity to quantization and generation schemes: The alignment between the model's structural complexity and the nature of the ground truth spatial queries suggests that, for extraction and classification tasks of low semantic complexity but high geographic specificity, compact models maximize inference efficiency without suffering the degradation effects that sometimes arise when fine-tuning oversized architectures on scarce data[31].

To examine more closely the nature of the inconsistencies corrected by the designed pipeline, Figure 1 presents the confusion matrices derived from the classification process, which allow the distribution of correct classifications and false alarms in each category to be compared visually.

A detailed examination of Figure 1 reveals a severe structural limitation in the behavior of Geocoder-Base, characterized by a marked tendency to underestimate or discard potentially valid records. Specifically, the traditional algorithm misclassified a total of 99 addresses belonging to the "Malla Vial" typology and 25 corresponding to "Manzana y Casa", incorrectly assigning them to the "Desconocido o Incompleto" category.

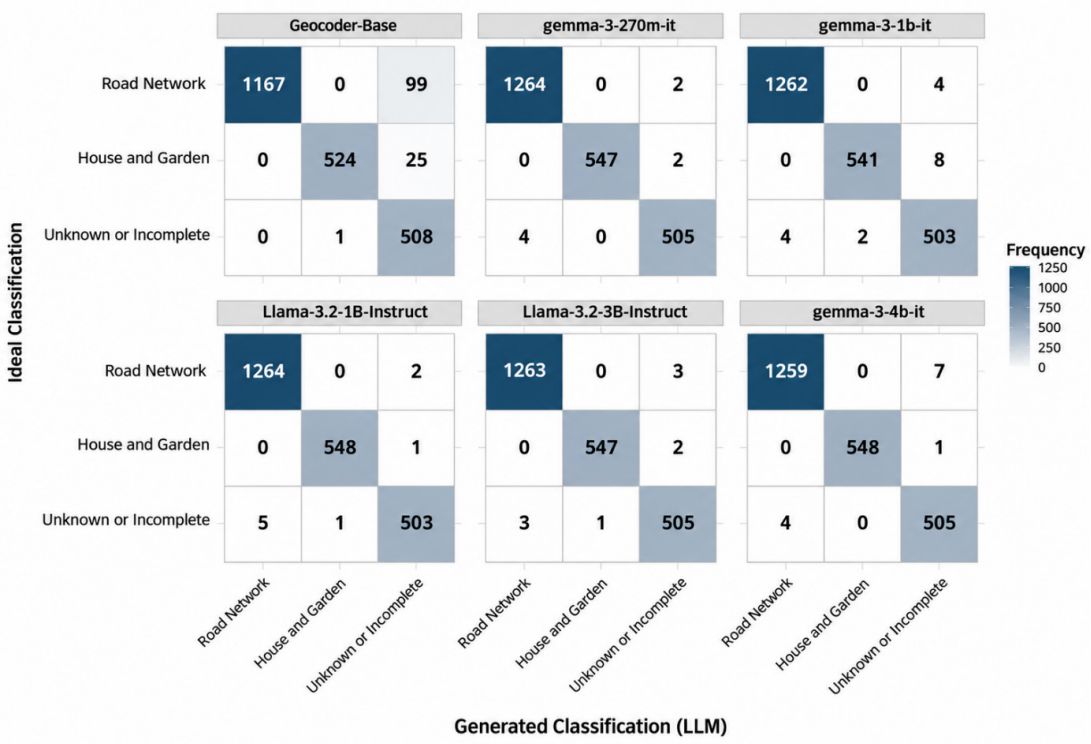

Fig. 1. *Confusion Matrix and Model Evaluation Metrics.*
Source: Own elaboration

By contrast, the implementation of the hybrid pipeline mitigated this classification bias almost entirely. The optimized architectures, such as Llama-3.2-3B-Instruct and Gemma-3-270m-it, drastically reduced these false rejections, restricting them to a marginal range of between 2 and 3 records. This substantial increase in analytical precision made it possible to recover critical spatial information that, under conventional deterministic logic, would have been regarded as null data loss.

Apart from categorical classification, the actual effectiveness of the hybrid system depends on its ability to accurately restructure the original text string in accordance with the official DANE standard[3], a step that is the indispensable requirement for definitive georeferencing. To evaluate this behavior, Figure 2 presents a comparative view of the success rate achieved in direct textual matching throughout the entire computational workflow.

Validation of Textual Normalization (End to End)

The findings consolidated in Figure 2 provide solid empirical support for the relevance and robustness of the proposed architecture. The traditional engine in isolation showed an operational limitation, reaching a maximum standardization ceiling of 84.08 %, whereas the hybrid workflow integrated into YACHAY-SIG correctly structured the information in the vast majority of highly complex records. Under this strict textual accuracy evaluation criterion, the pipeline coupled with the Llama-3.2-3B-Instruct architecture recorded the highest performance, reaching 99.40 % effectiveness in direct matching.

This evolution in the indicators confirms that the cascading processing strategy serves a dual purpose of high technical relevance: first, it acts as a safeguard mechanism for the computational infrastructure by restricting the activation of the large language model (LLM) exclusively to the failure scenarios of the Geocoder-Base; second, it definitively resolves the technological bottleneck caused by the extreme lexical variability and the use of informal descriptions in the oncology registry, guaranteeing data integrity for subsequent spatial analysis.

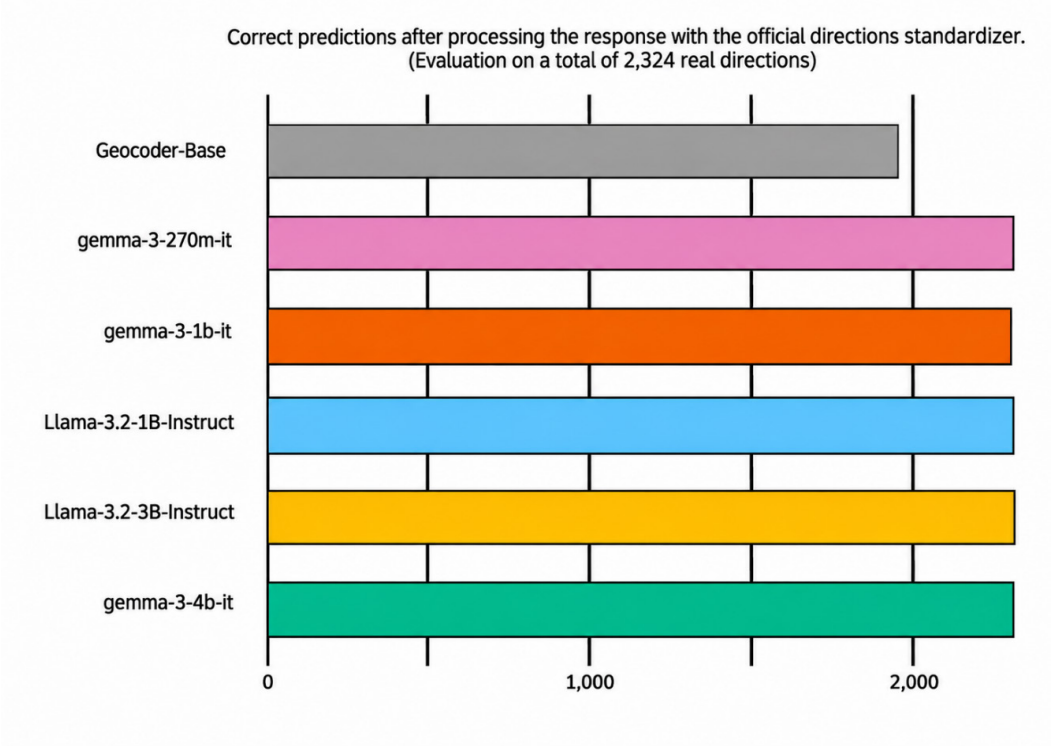

Fig. 2. *Accuracy rate in the textual normalization of addresses.*
Source: Own elaboration

Qualitative analysis: Success cases and persistent errors

To examine in greater depth the strengths and operational limitations of the proposed hybrid system, a qualitative analysis of the inferences generated on the test dataset was carried out. Table 4 illustrates representative examples of successfully corrected normalizations alongside scenarios in which discrepancies or failures in the final spatial assignment persisted.

Table 4: Representative examples of successful normalizations and persistent errors identified in the hybrid geocoding system

Case Type	Input Text (Original Query)	System Result (Hybrid Pipeline)
Success Case 1 (Local abbreviation / Toponymy)	Cra 24 con Cl 18 sur, sector de Pandiaco	Carrera 24 # Calle 18 Sur, San Juan de Pasto
Success Case 2 (Typographical error correction)	Av los Estudiantes in front of the udenar	Students' Avenue, University of Nariño
Persistent Error 1 (Homonym ambiguity)	Calle Principal, vereda Las Cochas.	Imprecise assignment of rural road segment
Persistent Error 2 (Severe fragmentation)	Diagonal 5 este # 12-bis (behind the store)	Failure in sub-address normalization

Source: Own elaboration

As Table 4 shows, in the success cases the fine-tuning strategy using LoRA allows the model to resolve complex linguistic ambiguities, non-standard cadastral abbreviations, and colloquial toponymic variations typical of the study region, which failed systematically in the baseline (Geocoder-Base). However, persistent errors are concentrated primarily in edge cases characterized by a high scarcity of adjacent spatial context, addresses with severely fragmented nomenclature, or the presence of multiple road homonyms in peripheral rural areas, where the model tends to overgeneralize or to require additional topological constraints in PostGIS.

Geospatial Impact, Integration into YACHAY-SIG and Cartographic Limitations

Once the effectiveness of the hybrid workflow in the field of textual normalization had been corroborated, the analysis was directed toward evaluating its direct cartographic impact on the RPCMP database. For this phase of operational deployment and validation in a real environment, the integrated system processed the global sample of 15,645 records analyzed in the initial quality diagnosis, enabling a direct methodological contrast between the performance of the proposed model and the deterministic baseline.

It is important to note that, unlike pure natural language processing (NLP) metrics, the definitive indicator of success for the geographic information system was defined as the geospatial recovery rate (Geocoding Match Rate). This metric strictly quantifies the proportion of clinical records that were successfully intersected with the spatial database, thus obtaining a valid geographic coordinate under the WGS84 reference system.

Table 5: Increase in the geospatial geocoding rate after the hybrid integration.

Processing phase	Rate	Improvement
Baseline (traditional only)	33.02%	-
Hybrid strategy (traditional + LLM)	55.13%	+66.9%

Source: Own elaboration

As broken down in Table 5, implementing the hybrid cascade system produced a significant increase in the spatial visibility of the epidemiological data. The proposed pipeline successfully geocoded a total of 8,625 records, in contrast with the 5,167 locations obtained through the traditional baseline.

This increase represents an operational transition from an initial coverage of 33.02 % to 55.13 % effectiveness, equivalent to a relative improvement of 66.9 % in the capacity to recover oncological cases within the territory. These indicators confirm the viability of the architecture for mitigating the bias caused by the loss of informal data in public health registry systems.

Semantic Gap versus Spatial Gap

The findings presented reveal a critical phenomenon of a methodological nature: although it was empirically demonstrated that the hybrid pipeline can textually normalize 99.40 % of complex addresses, the final cartographic success rate stabilized at 55.13 %. This mathematical divergence corroborates that the algorithm successfully resolved the so-called linguistic barrier; consequently, the remaining geocoding failures are no longer attributable to deficiencies in the software or the language model.

Instead, this behavior responds to a strict limitation of the spatial data infrastructure and the base cartography of the municipality. The addresses, despite having been correctly structured by the AI in accordance with the regulations, correspond to informal urban expansion areas, peripheral settlements or rural sectors that still lack geometric representation and road indexing in the official cartographic databases.

Spatial Validation and Visualization

To guarantee the integrity and territorial reliability of the recovered data, a geometric validation protocol was implemented using "point-in-polygon" topological tests on the official vector layers (shapefiles) of DANE[3]. This procedure made it possible to verify that 100 % of the recovered medical records are located strictly within the political-administrative boundaries of the municipality of Pasto, eliminating the presence of "orphan points" or erroneous coordinates outside the study area.

The definitive integration of these functionalities into the YACHAY-SIG module gives the system the capacity to identify clusters of oncological incidence and spatial distribution patterns in highly vulnerable sectors that previously lacked technical visibility. Figures 3 and 4 illustrate this distribution through point density maps and heat maps (kernel density), consolidating the platform as a technological and strategic input of high value for informed decision-making in public health and epidemiology[4].

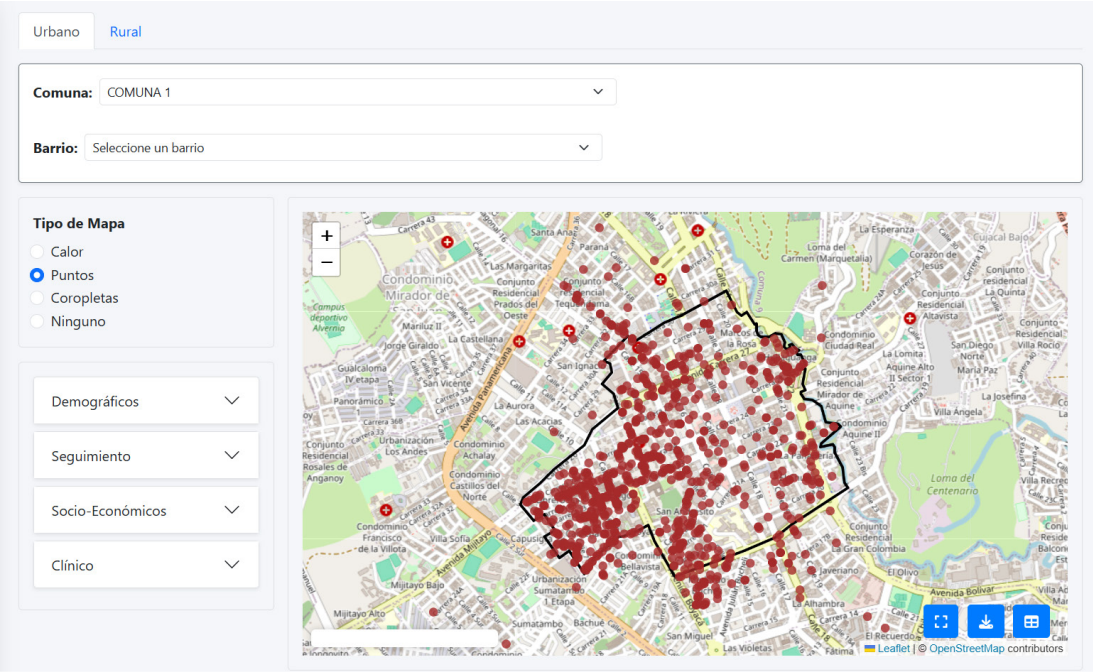

Fig. 3. *Point map of the distribution of cancer cases in Pasto.*
Source: Own elaboration

Discussion of Results and Limitations

The precision achieved in this study positions the YACHAY-SIG platform as a competitive alternative in intelligent geocoding for environments with data limitations. Comparing natural language processing with cartographic validation shows that the hybrid architecture optimizes computational performance and helps mitigate the impact of geospatial information bias in the cancer registry evaluated.

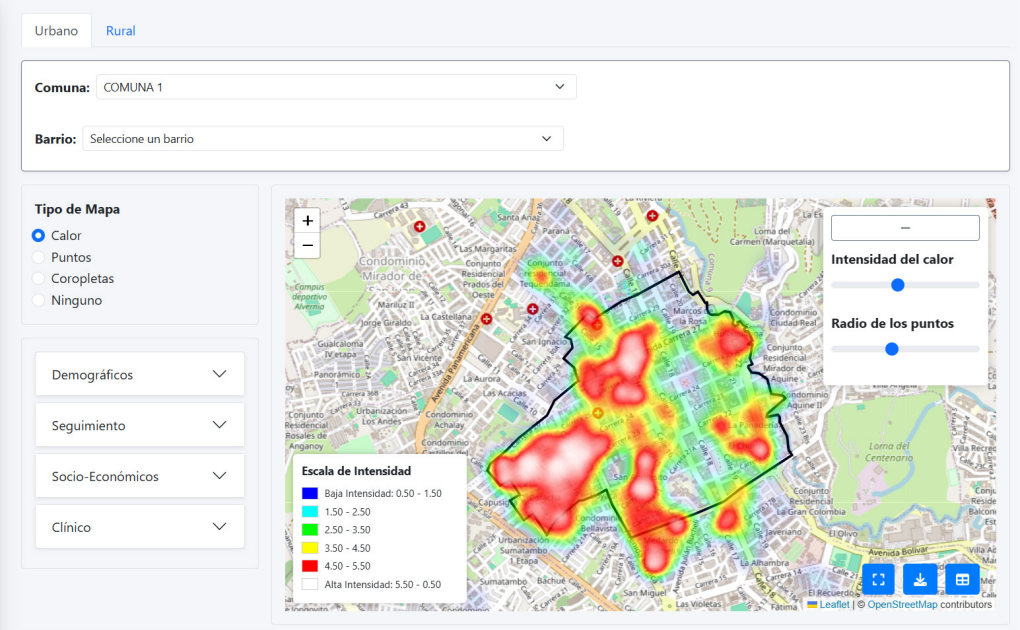

Fig. 4. *Heat map of the distribution of cancer cases in Pasto.*
Source: Own elaboration

In agreement with Guermazi et al. [13] in the logistics field, the results corroborate that neural architectures outperform traditional string matching algorithms when faced with noisy data. While the baseline deterministic geocoder reached an operational ceiling of 33.02 %, integrating the pipeline with the LLM agent (Llama-3.2-3B-Instruct) raised semantic standardization accuracy to 99.40 %. This made it possible to scale the overall geocoding rate to 55.13 %, a relative improvement of 66.9 % in data recovery (Table 5). These figures support the usefulness of moving from rigid systems toward models with semantic flexibility in this domain.

Likewise, the implementation of parametrically efficient techniques such as LoRA [24] demonstrates the technical feasibility of adapting general-purpose language models to specialized domains under computational resource constraints. In the context of Pasto, this fine-tuning facilitated the processing of "neighborhood, block and lot" type nomenclatures, a technical challenge in which conventional methods tend to have limitations in interpreting the lexical variability and urban context of the region.

From a public health perspective, authors such as McDonald et al. [11] argue that the quality of geographic data in oncology frequently depends on manual corrections to mitigate spatial biases. In this sense, the hybrid strategy of YACHAY-SIG shows the capacity to automate the interpretation of addresses with a residual error rate in normalization of 0.60 %. By processing the set of records analyzed, the system contributes to reducing the territorial omission of sectors that present difficulties in traditional epidemiological maps.

An analytical finding of value for urban planning is that the limitation of the final geocoding rate (55.13 %) does not lie in the model's textual comprehension capacity (99.40 %), but rather in the absence of specific geometries and roads in the official reference cartography. This points to a restriction in the available spatial data infrastructure (SDI), marking a point of attention for future cadastral updates.

Despite the favorable performance obtained by the hybrid system, some methodological and technical limitations should be noted, as they open the way for future lines of research. First, the results show a strong dependence on the quality and completeness of the official spatial data infrastructure (SDI) and of the available municipal cartography. As observed in the geocoding analysis, the lack of updated geometries or of specific road segments in cadastral sources limits the potential for final spatial recovery, regardless of the high metric precision achieved in the semantic standardization of the text.

Second, the scope of the experimental validation was confined to the specific domain and territorial context of the study region, without incorporating tests on external datasets or on other municipalities with heterogeneous toponymic structures.

Although the fine-tuning strategy using LoRA demonstrated local robustness, generalizing the architecture to other regions will require validation with additional heterogeneous corpora in order to rule out possible adaptation biases to the lexical particularities of the evaluated area.

Finally, the integration of these components through open-source tools such as PostgreSQL and Metabase illustrates a functional operational scheme for the YACHAY project. This cascading architecture model offers a useful methodological approach to be analyzed and adapted by other population registries that face similar challenges in the standardization of their household nomenclature.

CONCLUSIONS AND FUTURE WORK

This research shows that the integration of Large Language Models specialized through the Low-Rank Adaptation (LoRA) technique constitutes an effective alternative for the data normalization of heterogeneous geographic data in the evaluated domain. By overcoming the limitations of conventional deterministic methods, which recorded a base coverage of barely 33.02 %, the hybrid pipeline supported by the fine-tuned architecture reached a textual accuracy of 99.40 % in the interpretation of complex nomenclatures, recovering epidemiological information that previously required manual intervention.

The implementation of a hybrid cascading processing strategy proved to be an efficient approach for handling the records. By using the traditional geocoder as the first line of processing and activating the intelligent model for ambiguous records, the use of computational resources was optimized. This synergy made it possible to raise the geospatial recovery rate to 55.13 % (a relative increase of 66.9 %), indicating that the model contributed to resolving the semantic barrier. This finding shows that a relevant limiting factor for achieving complete georeferencing lies in the coverage and updating of the official cartography available in informal expansion zones and rural areas of the municipality of Pasto.

Finally, this work confirms the methodological usefulness of using open-weight models and their deployment on open-source tools such as PostgreSQL and PostGIS, offering a replicable architecture to be analyzed by other population registries or epidemiological surveillance systems that face similar challenges in the standardization of their household nomenclature.

As future work, this research points toward the operational implementation of the developed system. As an immediate step, we contemplate deploying the YACHAY-SIG platform in the environment of the Population-based Cancer Registry of the Municipality of Pasto (RPCMP). Once in operation, the normalization and automatic geocoding processes described here are expected to contribute to generating epidemiological cartography that supports the identification of spatial patterns and reduces institutional management times.

The research also plans to incorporate reverse geocoding techniques into the processing pipeline to evaluate the topological coherence between the coordinates obtained and the original textual descriptions, as a complementary mechanism for quality control and spatial auditing.

Finally, we plan to explore technology transfer and the adaptation of the models to other regional contexts in Colombia, in order to evaluate their performance under new toponymic structures and to contribute to strengthening spatial information systems in public health.

REFERENCES

[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.

FUNDING

This project is funded with resources from the Research System of the University of Nariño.

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Silvio Ricardo Timarán Pereira, Jonathan David Viveros Córdoba.

Writing–review and editing: Silvio Ricardo Timarán Pereira, Jonathan David Viveros Córdoba.