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

PhD. Silvio Ricardo Timarán Pereira[1], **Ing. Jonathan David Viveros Córdoba**[1]

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

Correspondence: ritimar@udenar.edu.co

Received: april 09, 2026. **Revised:** june 22, 2026. **Accepted:** july 24, 2026. **Published:** july 28, 2026.

How to cite: S. R. Timarán Pereira and J. D. Viveros Córdoba, "A Cascading Hybrid Strategy for Address Normalization using Fine-Tuned LLMs to Strengthen Cancer Registry Geospatial Analysis", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 48, pp. 166–180, jul. 2026, doi: 10.24054/rcta.v2i48.4512

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.

1. INTRODUCTION

Globally, public oncology demands epidemiological surveillance systems with analytical capabilities that transcend mere incidence registries, prioritizing instead the comprehension of the spatial distribution of pathology. Within this scenario, the Population-based Cancer Registry of the Municipality of Pasto (Registro Poblacional de Cáncer del Municipio de Pasto - RPCMP) constitutes the central node for the analysis of the neoplastic burden in the region. However, transforming these data into strategic assets for clinical and administrative decision-making imperatively requires high-precision georeferencing mechanisms that enable reliable multi-scale visualization.

In response to this technical necessity, YACHAY-GIS [1] has been implemented an intelligent geographic information infrastructure designed to robustly enhance the analytical capacity of the RPCMP. This platform is underpinned by a relational data architecture managed in PostgreSQL with the PostGIS geospatial extension, ensuring optimal efficiency in the storage and processing of clinical and geographic variables. The system enables the disaggregation of prevalence and incidence across micro-local territorial units (neighborhoods, communes, townships, and rural sectors), facilitating cartographic synthesis through heatmaps and dynamic choropleth maps [2].

Nonetheless, the operational integrity of YACHAY-GIS is intrinsically linked to the quality of the source data (input data). Patient location is derived from postal addresses captured in healthcare environments, which exhibit high entropy and structural variability. Previous diagnostics confirm that ambiguity in nomenclature hinders alignment with the National Geostatistical Framework of the National Administrative Department of Statistics

(Departamento Administrativo Nacional de Estadística - DANE) [3], the official reference cartographic baseline [4]. This issue worsens in the city of Pasto, where the coexistence of official nomenclatures with informal descriptors based on "blocks and lots" (manzanas y lotes) generates a significant semantic interoperability gap [5].

In recent years, the evolution of geospatial artificial intelligence (GeoAI) has transformed how unstructured spatial data is processed and normalized [6]. Traditionally, geographic information systems relied on rigid rules and heuristic methods that exhibited severe limitations when dealing with imprecise, colloquial, or incomplete location descriptions [7]. However, to overcome the limitations inherent in character-based matching and traditional schema approaches, recent research has increasingly incorporated Natural Language Processing (NLP) techniques based on Transformer architectures [8]. Specifically, studies have shown that models such as BERT and RoBERTa significantly improve geo-entity extraction and address matching by capturing the contextual semantics of text [8], [9].

More recently, the emergence of Large Language Models (LLMs) has introduced a new paradigm in intelligent geocoding, enabling the transformation of coordinate prediction into advanced generative tasks and resolving complex locality descriptions with relative spatial relationships [10]. These technologies provide a fundamental pathway to bridge the semantic gap in complex domains such as health and population registries [6], [10].

Historically, geocoding within non-standardized nomenclature environments has been addressed through various methodologies. At the regional level, deterministic engines based on string-matching algorithms optimized for the local road network have been developed [5]. Although these systems provide a solid foundation, their performance decreases exponentially when faced with unstructured records, which are frequent in peripheral and rural areas. Concurrently, international scientific literature indicates that the manual processing of these coordinates is economically unfeasible for large data volumes [11], which has catalyzed the adoption of deep learning models and Transformer architectures (such as BERT and RoBERTa) for the resolution of textual ambiguities [12], [13].

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

Despite the methodological advances described, a significant scientific gap persists in the automated normalization of high-entropy addresses [16]. Traditional approaches, grounded in deterministic algorithms and heuristic string-matching rules, operate under the assumption of a rigid urban nomenclature, exhibiting operational collapse when faced with severe typographic errors, informal abbreviations, or the absence of essential normative components [17]. With the adoption of classical Transformer architectures, previous research managed to partially mitigate 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 (e.g., descriptions such as "behind the main park, vacant lot") [18].

The LLM-based strategy addresses this gap through a paradigm shift by transforming normalization into a structured generative task capable of reasoning about spatial ambiguity [10]. Nevertheless, 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 low computational cost techniques (LoRA) [13], a fundamental shortcoming that the present work aims to resolve.

Compared to other hybrid proposals reported in the literature which generally limit themselves to sequentially connecting black-box commercial geocoding services with generic text-cleaning heuristics [19], the differentiating contribution of this research lies in the design of a tightly-coupled framework optimized for regional contexts with limited resources. Instead of relying on proprietary massive models, the proposed strategy integrates a local deterministic engine in PostgreSQL/PostGIS with an open-weight language agent specialized through parameter-efficient adaptation (LoRA) [20]. This approach not only guarantees the

sovereignty and privacy of sensitive epidemiological data from 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 recovery rate [19], [20].

This article presents the implementation of a hybrid strategy that amalgamates 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, providing a solid scientific foundation for oncological analysis with a territorial approach.

2. MATERIALS AND METHODS

2.1. Materials

The design and implementation of the address normalization module, as well as its subsequent integration into the YACHAY-GIS platform, were structured around three essential components: official data repositories, state-of-the-art Large Language Models (LLMs), and a high-capacity dedicated computing infrastructure for processing.

2.1.1. Data Sources and Cartographic References

As the primary input, historical records from the Population-based Cancer Registry of the Municipality of Pasto (RPCMP) were utilized. These records consolidate the clinical and sociodemographic data of the population diagnosed within the region.

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

Table 1: *RPCMP Dataset Partition*

Category / Subset	No. Reg.	Exclusion / Inclusion Criteria	Composition / Distribution
Initial Raw Records	15,645	Total extraction from the RPCMP database	Totality of raw oncological notifications
Excluded Records (Data Cleaning)	(153)	Null coordinates or 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 clean data integrated into the hybrid pipeline
Training Subset	10,844 (70%)	Assigned for fine-tuning via LoRA.	Stratified distribution by communes and severity
Validation Subset	2,324 (15%)	Used for hyperparameter monitoring and early stopping	Independent representative sample.
Test Subset	2,324 (15%)	Reserved for blind evaluation of final performance.	Complex and high-entropy rural cases

Source: *own elaboration*

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

2.1.1.1 National Geostatistical Framework (MGN): Provided by DANE [3], this resource allowed for the establishment of the official politico-administrative structure, ranging from the division by neighborhoods and urban sectors to the rural organization into townships (corregimientos) and rural sectors (veredas).

2.1.1.2 Pasto Geocoding Engine: The open-source tool developed by Timarán et al. [5] was integrated. This engine employs string-matching algorithms to convert physical addresses into geographic coordinates and is essential as it is specifically

calibrated for the road topology and nomenclature unique to San Juan de Pasto.

2.1.2. Large Language Models (LLMs)

For the experimentation and fine-tuning phases, five variants belonging to two Open Weights model families were selected. These models were chosen based on the need to balance processing power with operational efficiency in hardware architectures with constrained capacities:

2.1.2.1 Gemma 3 IT Series (270M, 1B, and 4B): Developed by the Gemma Team [21], these models represent the state of the art in lightweight architectures. They are characterized by optimized logical reasoning and adherence to complex instructions while maintaining reduced power and memory consumption.

2.1.2.2 Llama 3.2 Series (1B-Instruct and 3B-Instruct): These compact versions from Meta AI [15] were implemented, having been refined for high-performance deployments. They stand out for their advanced semantic comprehension and for ensuring low latency in limited computing environments.

2.1.3. Software Ecosystem and Development Tools

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

2.1.3.1 Spatial Data Management: PostgreSQL was adopted as the relational database management system, complemented by the PostGIS extension for the administration, storage, and execution of geometry queries referenced in the WGS84 coordinate system.

2.1.3.2 Programming Environment and Dependencies: System logic development was conducted in Python [22], utilizing the *uv* package manager to ensure efficient and isolated dependency management.

2.1.3.3 Training and Fine-Tuning: The training and optimization phases were executed through the combined use of LLaMA Factory [23] and Unsloth. The adaptation of the models was based on efficient parameterization techniques LoRA (Low-Rank Adaptation) [24], executed on high-performance graphical infrastructure with CUDA support.

2.1.3.4 Result Visualization: The spatial representation of information was integrated into the YACHAY-GIS web interface a platform developed on modern web architectures oriented toward the generation and visual analysis of dynamic heatmaps and choropleth maps.

2.1.4. Hardware Infrastructure

The training and fine-tuning process of the models was supported by a high-performance workstation equipped with an NVIDIA A100 Tensor Core GPU (with 40 GB of VRAM) supporting CUDA technology. The computational workflow was developed on the Ubuntu 22.04 LTS operating system, utilizing Python 3.10.12, PyTorch 2.3.0 [22], and the Hugging Face Transformers library [25], employing LLaMA-Factory as the primary interface [23].

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

2.2 Methods

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

The methodological process was executed through the following interconnected phases:

2.2.1 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. Through the implementation of Natural Language Processing (NLP) techniques, inconsistencies in the records were systematically analyzed, detecting critical error patterns 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 fundamental to defining the specific training parameters and requirements, thereby guiding the fine-tuning of the language models toward resolving the lexical and structural particularities identified in the local data.

2.2.2 Corpus Construction for Fine-Tuning

With the objective of specializing the Large Language Model (LLM), a structured training dataset was developed through two fundamental processes:

2.2.2.1 Manual Labeling (Ground Truth): A comprehensive normalization procedure was carried out, in which domain experts manually analyzed and standardized each record in accordance with official address regulations. This process established the "ground truth," which is essential for ensuring the quality of the model's learning.

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

2.2.3 Fine-Tuning Configuration Using LoRA

The core component of the methodology lay in the implementation of the Low-Rank Adaptation (LoRA) technique [24]. Instead of executing a full retraining of the parameters in 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:

2.2.3.1 Resource Optimization: It allowed for a drastic reduction in VRAM memory consumption by freezing the original weights and exclusively updating the adaptation parameters.

2.2.3.2 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.

2.2.3.3 Computational Efficiency: The training was conducted in bfloat16 (bf16) precision format, optimized through the use of Gradient Checkpointing and Flash Attention.

At an operational level, 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 a gradient accumulation of 4 steps) to optimize stability in weight updates. 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 the minimization of the validation loss with a patience margin of 2 epochs to prevent overfitting [28]. Regarding the specific LoRA parameterization, a rank factor (r) of 16, a scaling factor (α) of 32, and a dropout rate of 0.05 applied over the attention projections were established [23].

2.2.4 Implementation of the Hybrid Geocoding Algorithm

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

2.2.4.1 Phase 0: Textual Preprocessing and Cleaning. Prior to the geospatial evaluation, the original text strings are subjected to a structured cleaning module. Through the use of regular expressions, the algorithm standardizes capitalization, eliminates anomalous punctuation, and executes a systematic separation of 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.

2.2.4.2 Phase 1: Enhanced Deterministic Engine (Proximity Search). At this stage, the process employs a base geocoder [5], which was strengthened through advanced heuristics and increased syntactic flexibility. Unlike the original strict-order approach, the new algorithm allows for the interpretation of inverted or partial nomenclatures. Furthermore, the exact matching 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 within the road network, the system

evaluates adjacent locations by applying penalty coefficients for numerical distances or variations in alphabetic suffixes. This allows for the assignment of the closest possible location coordinate, provided that a predefined metric confidence threshold is met.

2.2.4.3 Phase 2: Intelligent Retrieval via LLM. As a final processing stage, records presenting severe ambiguities, extreme colloquial references, or those resulting in a total failure of the deterministic engine are diverted to the previously fine-tuned Large 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 according to official DANE standards.

Once normalized, the string is re-injected 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 due to their low quality.

2.2.5 Evaluation Protocol and Success Metrics

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.

2.2.5.1 Linguistic Evaluation (Normalization Quality). To determine the technical proficiency of the fine-tuned models (Gemma 3 and Llama 3.2), the accuracy metric was used. The calculation was performed through a direct contrast between the addresses normalized by the model in the test dataset and the ground truth established by expert manual labeling. This evaluation allowed for measuring the models' capacity to abstract textual noise, correct spelling errors, and restructure the information under the official nomenclature standards defined by DANE [3].

2.2.5.2 Effectiveness Evaluation (Match Rate). Since ultimate spatial accuracy is conditioned by mapping coverage and the deterministic engine, the primary metric for the hybrid system was defined as the Geocoding Match Rate.

This indicator represents the percentage of addresses or geographic records that the system successfully locates and correctly associates with their spatial coordinates or corresponding reference points within the official geographic database. Operationally, a high value indicates that the model is capable of processing and structuring imprecise or incomplete location descriptions provided by citizens, successfully translating them into digital maps without significant information loss.

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

3. RESULTS AND DISCUSION

The analysis of the implemented strategy corroborated the effectiveness of Large Language Models (LLMs) as tools for retrieving geospatial data essential for the RPCMP. The following sections detail the findings, structured into the quality diagnosis, the performance of the fine-tuning process, and the impact achieved on the final georeferencing.

3.1 Data Quality Diagnosis in the RPCMP

In order to establish a baseline and determine the technical complexity of the cancer records in the municipality of Pasto, an initial diagnosis was conducted using the open-source geocoder developed by Timarán et al. [5]. Although this system represents a regional technological pillar, its deterministic architecture is designed for predictable road structures, which limits its performance when faced 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 frequently exhibit typographical errors and grammatical omissions, non-standardized abbreviations, and colloquial descriptions that are incompatible with conventional string-matching algorithms.

A key methodological factor was the processing of records under the two nomenclature models that coexist in the city: "road network (malla vial)" and "neighborhood-block (barrio-manzana)" A comprehensive protocol was applied where each

data point was evaluated by both functions, validating the result if either one successfully achieved coordinate assignment.

Results from a sample of 15,645 records highlighted the magnitude 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, thus being excluded from the initial spatial analysis. This considerable information gap, summarized in Table 2, justifies the implementation of LLMs with semantic flexibility to normalize and recover vital data for epidemiological surveillance.

Table 2: *Summary of Address Diagnosis*

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

Source: *own elaboration*

3.2 Performance of the Hybrid Strategy and System Validation

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

3.2.1 Evaluation Dataset Configuration

o 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. These instances were manually normalized and categorized by experts, 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).

It is important to emphasize that the analytical evaluation presented in this section was strictly performed on the test set (n = 2,324), which contains the records with the most critical levels of lexical ambiguity. This allowed for a direct comparison between the system's predictions and the standardized ground truth.

With the purpose of ruling out any type of bias and ensuring an objective performance evaluation, the test set was kept completely isolated and independent throughout the entire development cycle [29]. This set did not intervene in the construction phase nor in the debugging of the ground truth reference set, and its instances were not exposed in any way during the fine-tuning process via LoRA or in the validation-based hyperparameter selection. Thus, the final evaluation rigorously reflected the models' generalization capability on completely unseen data.

3.2.2 Effectiveness in Structural Classification

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

Based on the consolidated data in Table 3, it is evident that the implementation of the hybrid workflow consistently outperforms the baseline. The integration of the Gemma-3-270m-it model is especially noteworthy, as its inclusion raises the F1-score from 0.942 (recorded by the Base-Geocoder) to an optimal 0.996. This performance is particularly significant considering the computational efficiency of this architecture, which achieves critical levels of precision despite its reduced parameter count.

Table 3: *Comparative 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.*

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

The findings indicate that geographic retrieval precision is strongly coupled to accuracy in resolving local toponymic entities. In particular, the Gemma-3-270m-it model provided the highest effective geographic retrieval within the hybrid pipeline, achieving a correct spatial assignment rate of 99.6% on the test set. This behavior is explained by the fact that its lower degree of semantic overgeneralization prevents undesired alterations in the syntax of specific addresses and points of interest in the study region. On the other hand, larger architectures (Gemma-3-4b-it and Llama-3.2-3B-Instruct), while demonstrating fluent performance in general linguistic tasks, showed a slight degradation in final geocoding due to occasional hallucinations or over-normalization 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 to considerably larger architectures can be explained by three main technical hypotheses in the context of specialized fine-tuning:

1. Benign overfitting and local optimization: Models with fewer parameters possess a more restricted search space, which, under a low-rank adaptation (LoRA) strategy [24] with a bounded domain-specific corpus, facilitates rapid convergence and highly efficient coupling to local nomenclature and toponymy. In contrast, larger-scale models possess an oversized representation capacity for the provided training data volume, which can induce suboptimal optimization or interference with generic pre-trained weights [30].

2. *Adaptation dynamics in LoRA and base weight capacity:* In lightweight architectures, updating the low-rank matrices proportionally alters a larger effective proportion of the model's behavior relative to its original latent space. Conversely, 3B and 4B parameter models possess much deeper attention structures, which may require a larger volume of reinforcement data or a different hyperparameter scheme for adaptation to redefine 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 spatial queries in the ground truth suggests that, for extraction and classification tasks of low semantic complexity but high geographic specificity, compact models maximize inference efficiency without suffering from the degradation effects that sometimes manifest when applying fine-tuning to oversized architectures with scarce data [31].

To further examine the nature of the inconsistencies corrected by the designed pipeline, Figure 1 presents the confusion matrices derived from the classification process. These matrices allow for a visual contrast of the distribution of successes and false alarms across each category.

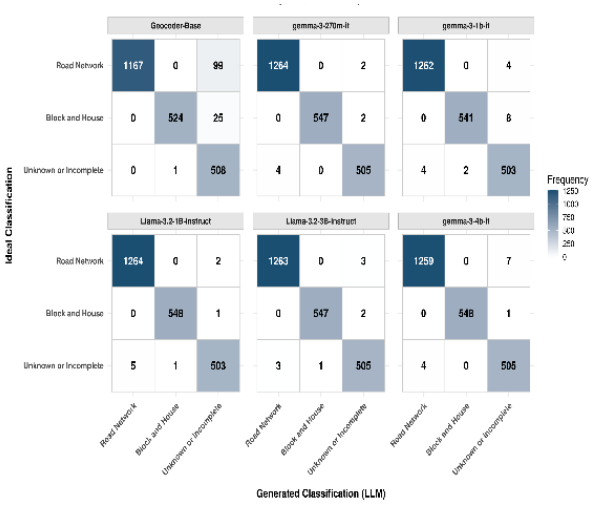

Fig. 1*. Confusion Matrix and Model Evaluation Metrics.*
Source: *own elaboration.*

A detailed examination of Figure 1 reveals a severe structural limitation in the performance of the Base-Geocoder, characterized by a marked tendency toward the underestimation or dismissal of potentially valid records. Specifically, the traditional algorithm misclassified a total of 99 addresses belonging to the "Road Network" typology and 25 corresponding to "Block and

House," incorrectly assigning them to the "Unknown or Incomplete" category.

In contrast, the implementation of the hybrid pipeline successfully mitigated this classification bias almost entirely. 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 allowed for the recovery of critical spatial information that, under conventional deterministic logic, would have been considered as null data loss.

3.2.3 Validation of Textual Normalization (End-to-End)

Beyond categorical classification, the actual effectiveness of the hybrid system is determined by its proficiency in accurately restructuring the original text string according to the official DANE standard [3], a step that constitutes the indispensable prerequisite for definitive georeferencing. To evaluate this performance, Figure 2 presents a comparative analysis of the success rate achieved in direct textual matching throughout the entire computational workflow.

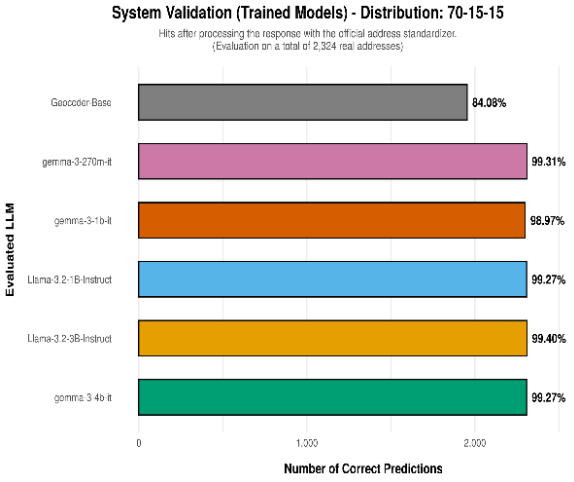

***Fig. 2**. Accuracy rate in textual address normalization.*
***Source:** own elaboration.*

The findings consolidated in Figure 2 provide solid empirical evidence regarding the relevance and robustness of the proposed architecture. While the isolated traditional engine exhibited an operational limitation, reaching a maximum standardization ceiling of 84.08%, the hybrid workflow integrated into YACHAY-GIS successfully structured the information in the vast majority of high-complexity records. Under this criterion of strict textual accuracy evaluation, the pipeline coupled with the Llama-3.2-3B-Instruct architecture recorded the highest performance, achieving 99.40% effectiveness in direct matching.

This evolution in the indicators confirms that the cascading processing strategy fulfills 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 failure scenarios of the Base-Geocoder; second, it definitively resolves the technological bottleneck caused by extreme lexical variability and the use of informal descriptions in the oncological registry, guaranteeing data integrity for subsequent spatial analysis.

3.2.4 *Qualitative Analysis: Success Cases and Persistent Errors*

With the purpose of delving deeper into the operational strengths and limitations of the proposed hybrid system, a qualitative analysis of the inferences generated on the test set was carried out. Table 4 illustrates representative examples of successfully corrected normalizations versus scenarios where 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 frente a la Udenar	Avenida de los Estudiantes, Universidad de Nariño
Persistent Error 1 (Ambiguity due to homonymy)	Calle Principal, vereda Las Cochas.	Imprecise assignment of rural road segment
Persistent Error 2 (Severe fragmentation)	Diagonal 5 este # 12-bis (atrás de la tienda)	Failure in sub-address normalization

***Source:** own elaboration.*

As observed in Table 4, in the success cases, the fine-tuning strategy via LoRA enables the model to resolve complex idiomatic ambiguities, non-standard cadastral abbreviations, and colloquial toponymic variations typical of the study region, which systematically failed in the baseline (Geocoder-Base). However, persistent errors are

primarily concentrated 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 require additional topological constraints in PostGIS.

3.3 Geospatial Impact, Integration into YACHAY-GIS, and Cartographic Limitations

Once the effectiveness of the hybrid workflow in the field of textual normalization was corroborated, the analysis shifted toward evaluating its direct cartographic impact on the RPCMP database. For this phase of operational deployment and real-world validation, the integrated system processed the global sample of 15,645 records analyzed in the initial quality diagnosis, allowing for 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 success indicator for the geographic information system was defined as the Geocoding Match Rate. This metric strictly quantifies the proportion of clinical records that were successfully intersected with the spatial database, thereby obtaining a valid geographic coordinate under the WGS84 reference system.

Table 5: *Increase in geospatial geocoding rate following hybrid integration*

Processing Phase	Rate	Improvement
Baseline (Traditional only)	33.02%	-
Hybrid Strategy (Traditional + LLM)	55.13%	+66.9%

Source: *own elaboration.*

As detailed in Table 5, the implementation of the cascading hybrid system produced a significant increase in the spatial visibility of the epidemiological data. The proposed pipeline successfully georeferenced a total of 8,625 records, in contrast to the 5,167 locations obtained using the traditional baseline.

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

3.3.1 Semantic Gap vs. Spatial Gap

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

Instead, this performance reflects a strict limitation of the spatial data infrastructure and the municipality's base mapping. Despite being correctly structured by the AI according to regulations, these addresses correspond to areas of informal urban expansion, peripheral settlements, or rural sectors that still lack geometric representation and road indexing in official cartographic databases.

3.3.2 Spatial Validation and Visualization

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

The final integration of these functionalities into the YACHAY-GIS module provides the system with the capability to identify oncological incidence clusters and spatial distribution patterns in high-vulnerability 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 high-value technological and strategic asset for informed decision-making in public health and epidemiology [4].

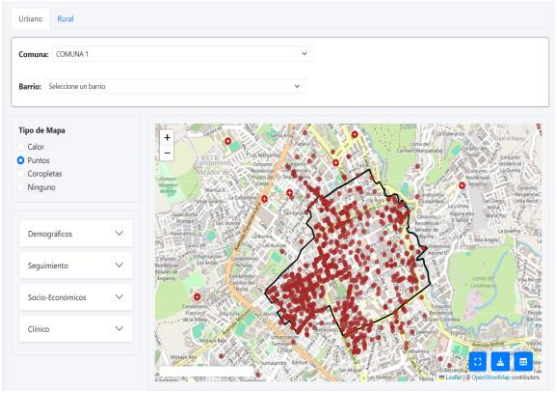

***Fig. 3**. Point distribution map of cancer cases in Pasto.*
***Source:** own elaboration.*

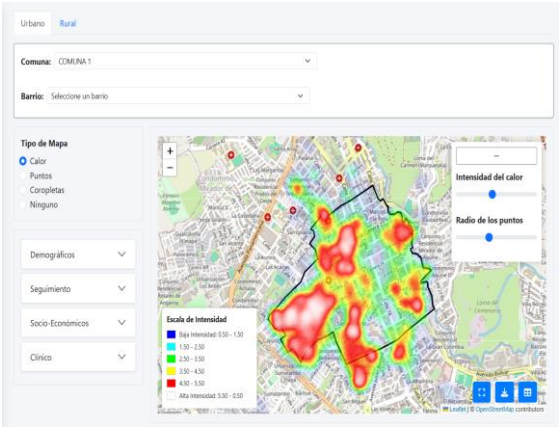

***Fig. 4**. Heat map of the distribution of cancer cases in Pasto..*
***Source:** own elaboration.*

3.4 Discussion of Results and Limitations

The precision achieved in this study positions the YACHAY-GIS platform as a cutting-edge solution for intelligent geocoding in developing environments. By contrasting natural language processing with cartographic validation, it becomes evident that the hybrid architecture not only optimizes computational performance but also definitively isolates the source of geospatial information bias within the oncological registry.

In accordance with the findings of Guermazi et al. [13] in the logistics domain, the results corroborate that neural architectures significantly outperform traditional string-matching algorithms when handling noisy data. While the base deterministic geocoder reached an operational ceiling of 33.02%, the integration of the pipeline with the LLM agent (Llama-3.2-3B-Instruct) raised the semantic standardization accuracy to 99.40%. This facilitated scaling the global geocoding rate to 55.13%, representing a 66.9% relative improvement in data recovery (Table 5). These figures validate the necessity of transitioning from rigid systems toward models with semantic flexibility.

Furthermore, the implementation of parameter-efficient fine-tuning techniques such as LoRA [24] demonstrates the feasibility of adapting general-purpose language models to specialized domains without incurring high computational costs. In the context of Pasto, this fine-tuning was the determining factor for correctly processing "neighborhood, block, and lot" (barrio, manzana, y lote) nomenclatures a technical challenge where commercial geocoders typically fail due to their inability to interpret the lexical variability and the informal 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 costly manual corrections to avoid spatial biases. However, the YACHAY-GIS hybrid strategy demonstrates that address interpretation can be automated with a residual algorithmic error rate of just 0.60%. By spatially projecting 8,625 records, the system mitigates the territorial invisibility of peripheral and rural populations that have historically been excluded from epidemiological maps.

A high-value finding for urban planning is that the limitation of the final geocoding rate (55.13%) does not lie in an inability to comprehend text (99.40%), but rather in the absence of polygons and roads in the official mapping. This highlights a gap in the municipality's Spatial Data Infrastructure (SDI), setting a precedent for future cadastral updates.

Despite the favorable performance achieved by the hybrid system, certain methodological and technical limitations must be noted that pave the way for future lines of research. First, the results reveal a strong dependency on the quality and completeness of the official spatial data infrastructure (SDI) and available municipal cartography. As observed in the geocoding analysis, the lack of updated geometries or specific road segments in cadastral sources limits the potential for final spatial retrieval, regardless of the high metric accuracy achieved in the semantic standardization of the text.

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

structures. Although the fine-tuning strategy via LoRA demonstrated local robustness, the generalization of the architecture to other regions will require validation with additional heterogeneous corpora to rule out potential adaptation biases toward the lexical particularities of the evaluated area.

Finally, the consolidation of these components into a technological sovereignty platform, based on open-source tools such as PostgreSQL and PostGIS, ensures the efficiency and scalability of the YACHAY project. This cascading architecture model serves as a replicable benchmark for other population registries in countries facing similar challenges in home nomenclature standardization.

4. CONCLUSIONS AND FUTURE WORK

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

The implementation of a cascaded hybrid processing strategy proved to be an efficient approach for handling records. By employing the traditional geocoder as the first line of processing and activating the intelligent model upon ambiguous records, the use of computational resources was optimized. This synergy allowed raising 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 evidences 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 ratifies the methodological usefulness of using open-weight models and 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 facing similar challenges in standardizing their residential nomenclatura.

As future work, the horizon of this research is directed towards the operational implementation of the developed system. As an immediate step, the deployment of the YACHAY-GIS platform within the environment of the Population-Based Cancer Registry of the Municipality of Pasto (RPCMP) is contemplated. With its operation, it is expected that the described automatic normalization and georeferencing processes will contribute to the generation of epidemiological cartography aimed at supporting the identification of spatial patterns and optimizing institutional management times.

Likewise, the research projects the incorporation of reverse geocoding techniques within the processing pipeline to evaluate the topological coherence between the obtained coordinates and the original textual descriptions, serving as a complementary quality control and spatial auditing mechanism.

Finally, plans are underway to explore technological transfer and the adaptation of the models in other regional contexts of Colombia, with the purpose of evaluating their performance against new toponymic structures and contributing to the strengthening of spatial information systems in public health.

Acknowledgments: This project is funded with resources from the Research System of the Universidad de Nariño.

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.