Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Where Exactly Are You? Disaster Response in India

Evaluating LLM-Enhanced Geocoding for Informal Locations in Low-Resource, Multilingual Contexts

Author: Adithya Vasisth (2762203)

1st supervisor: Dr. Anna Bon
2nd reader: Prof.Dr. Hans Akkermans

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

November 18, 2025

Abstract

Crisis mapping technologies assume everyone has formal addresses, reliable internet connectivity, and communicates in English. Most disaster-affected communities have none of these. When disasters strike, affected populations report emergencies through platforms promising democratized response. Yet a persistent bottleneck emerges: translating informal location descriptions into coordinates that response systems understand.

This research explores whether contemporary geocoding services and Large Language Model augmentation can address these challenges.

Using 117 authentic Bangalore flood addresses collected through an NGO's disaster relief work, we evaluated five geocoding services and tested whether LLM preprocessing could improve disaster response performance in low-resource, multilingual contexts that existing research overlooks.

Results expose stark inequality: commercial services achieved moderate accuracy (Google Maps 63%, OLA Maps 41%), while open-source alternatives failed catastrophically (2-5%). LLM augmentation improved some configurations but degraded others—a paradox revealing how proprietary systems handle informal addresses differently.

Technically, the technology works. Response times could drop from hours to seconds, potentially saving lives. But automation addresses one bottleneck while structural barriers—digital divides, training bias, power imbalances—remain. Whether deployment serves communities or optimizes extraction depends on who benefits, who decides, and whose knowledge counts.

Contents

1

Introduction

When disaster strikes, survival hinges on a deceptively simple question: where exactly are you?

In 2008, when I was 9 years old, I traveled through three Indian cities during what should have been an ordinary vacation with my family, spending one day in each. I didn't know that within 24 hours of leaving each city, a catastrophic flood would submerge the places I had just visited. The hotel room I stayed in one night was underwater within hours of my departure. Other tourists were stranded there and couldn't return home for a week. I watched news reports showing people struggling for food and supplies being airdropped by helicopters. These cities had been experiencing severe drought just days before water levels reached two stories high as rivers swelled beyond their banks. Had my travel been delayed by a few hours, I too would have been trapped. I only learned about this disaster after reaching home safely, the narrow escape feeling surreal in retrospect.

Figure 1.1: Hampi, a UNESCO World Heritage Site in Karnataka, submerged in Tungabhadra river waters during 2022 flooding. This was one of the three cities I visited during my 2008 vacation that experienced similar catastrophic flooding shortly after my departure(1).

That experience never left me. When I joined Robin Hood Army years later, those memories shaped how I understood what disaster response actually meant. For five years that I spent with them, I organized relief efforts at least three times yearly. I collected food and essential supplies for flood victims across India, posting on social media asking for help, gathering responses through word of mouth, organizing volunteers. I coordinated logistics to send materials to cities devastated by cyclones, particularly Odisha and Chennai. Sometimes I collected and delivered supplies myself. Other times I coordinated with different NGOs or disaster response agencies better positioned to reach specific locations. Once, I organized distribution in my own city when it experienced seasonal monsoon floods, managing volunteers on the ground and ensuring help reached those who needed it.

Figure 1.2: Disaster relief coordination with Robin Hood Army, Bangalore (2018-2019). Top left: Relief materials being dispatched by train to Jabalpur during flooding (August 24, 2019). Top right: Distribution of supplies to flood-affected families in Bangalore (August 19, 2019). Bottom left: Volunteers coordinating relief distribution (August 15, 2018). Bottom right: Robin Hood Army Bangalore chapter volunteers (June 2, 2019). Source: Personal collection.

Through this work, I learned that the challenge of disaster response extends far beyond collecting supplies. An incident that I coordinated particularly illuminates this. An elderly couple in Thanisandra, a locality in the rapidly urbanizing part of North Bangalore, had watched floodwater submerge their apartment's ground floor. They had run out of their medication. The underbridge leading to their locality lay flooded with rain water, cutting off vehicle access entirely. A neighbour with a working landline managed to call us for help, but the line was barely active. When I asked for their location, the response seemed straightforward: the apartment name and locality. The problem is that neither of these were unique. Dozens of apartments and localities share similar names across Bangalore's geography. The landline connection kept dropping. Each callback from a volunteer closer to their location meant starting over, asking the same questions, trying to extract landmarks that might distinguish this building from countless others. The help that should have reached them in a few hours took two whole days.

I constantly encountered this pattern. Messages would arrive via WhatsApp or Instagram: "My parents are stuck, they need help." Often these came from children living in different cities, anxious about elderly parents trapped in flooded neighbourhoods. I would begin the process of figuring out where people actually were. I learned to ask for a larger locality first, then landmarks, road names if any existed, building names or numbers when recorded. I would relay this information to another volunteer who lived closer to the affected area, passing along phone numbers so the local volunteer could call the family directly and navigate using their description. The process took around five minutes per address when the phone connectivity was strong and the communication remained clear. Often, neither condition is met during disasters.

Through coordinating these relief efforts, I came to understand how disasters unfold and what infrastructure actually survives. When floods hit, I watched failures cascade through systems that these communities rely upon. Power outages cut mobile networks first. Fallen trees severs broadband internet cables or power lines. What remained were some sporadic landline connections and desperate attempts to communicate through patchy WhatsApp messages during some brief windows of connectivity. I learnt that official disaster response agencies usually arrive in hours, but they need a few days to reach people due to infrastructure damage and severe weather conditions. Until then, neighbors become first responders. Organizations like mine scrambled to coordinate relief using whatever communication channels survived. Lives often really do depend on how quickly we could answer the question: where exactly are you?

Working with this voluntary organization in Bangalore revealed that the challenge extended beyond communication infrastructure alone. In interviews I conducted with other volunteers, a persistent pattern emerged. Rescue operations repeatedly encountered addresses like "second house after the bridge" or "behind the old temple." I saw how these descriptions carried precise meaning locally but remained a puzzle for geocoding services designed around formalized addressing conventions. This revealed a fundamental tension: rich contextual knowledge embedded in communities versus structured coordinate systems demanded by formal response infrastructure. This represents

a fundamental epistemic mismatch. Whose geographic knowledge gets formalized, digitized, and deemed "actionable"?

My research focuses specifically on addressing this problem in low-resource, multilingual contexts where commercial solutions fail most dramatically. While companies like Google optimize their geocoding services for English-speaking users with reliable infrastructure and formal addressing systems, disaster-affected communities in South Asia often have none of these advantages. This is the gap my work aims to fill.

The mismatch between local knowledge and formal systems isn't unique to the work I did in India. The pattern emerges wherever disasters strike and formal systems clash with local knowledge. Haiti's 2010 earthquake remains one of the starkest examples. In the chaos following the quake, a crisis mapping platform called Ushahidi became a central hub for coordinating information. Ushahidi had been developed in Kenya two years earlier specifically to crowdsource crisis reports from people on the ground. The platform collected text messages, social media posts, and direct submissions from Haitians trying to report emergencies, request help, or share information about damage. Over 100,000 crisis reports poured into the system. The sheer volume seemed promising. Finally, technology was enabling people to communicate their needs directly.

But here's what actually happened: only 3,584 of those reports ever got mapped to actual locations (2, 3, 4). That is only 5%. Over 96,000 reports containing potentially life-saving information sat in the system unusable because volunteers couldn't figure out where the reported addresses were. Coordination problems played a role. Verifying information proved difficult. But geocoding created a critical bottleneck. Remote volunteers working from different countries simply lacked the contextual knowledge to interpret location descriptions that made perfect sense to the locals but seemed impossibly vague to outsiders. "Trapped near collapsed church, water rising" may have contained vital intelligence if they knew Port-au-Prince's neighborhoods and which churches people used as reference points. For remote mappers staring at satellite images and street maps, it provided almost nothing actionable. Nepal's 2015 earthquake repeated the pattern. So did Pakistan's 2022 floods. Each time, communities generated crucial intelligence about their own crises. Each time, geographic knowledge gaps prevented that information from being used effectively.

These mismatches costs lives. Technology that should be used to coordinate rescue instead becomes a barrier determining who receives timely assistance. This gap is not just technical but structural, reflecting whose emergencies get prioritized by the systems we build.

South Asian contexts add specific complications. Indian addressing evolved through centuries of landmark-based wayfinding rather than Cartesian grids (5). Addresses like "palasi mohania village madarsa chowk rto jama masjid" conflate locality, landmark, and municipality without hierarchy, encoding social relationships rather than administrative boundaries (6). Regional variations multiply: "raste" (road) in Karnataka becomes "galli" in Delhi. Landmarks shift from temples to mosques to markets.

Bangalore itself exemplifies this complexity. Built over valleys, natural waterways, and former lakebeds, the city's unplanned expansion created neighbourhoods where addressing becomes almost

deliberately ambiguous. My own house has at least five different ways to describe its location, with five different versions recorded across official documents. Streets carry names in Kannada, English, or any other regional language. The British colonial heritage mingles with names honouring Indian and international personalities. Often, new localities emerge without government approval or sanction, appearing on no official maps yet housing thousands of families. When addressing itself is fluid, geocoding during a disaster becomes exponentially harder.

These addressing practices aren't failures. They are evolved systems adapted to local realities. Yet divergence from formalized standards creates systematic exclusion. OpenStreetMap coverage gaps persist in informal settlements where formal addresses vanish (7). Government standardization projects (India Post's PIN codes, recent addressing initiatives) fail to capture fine-grained distinctions meaningful to residents, particularly where land tenure ambiguity prevents formal assignment (8). Whose knowledge gets digitized determines who receives faster assistance. Communities with formalized addressing get precision. Others get delays. So what tools do responders and volunteers actually use when disasters strike, and why do they consistently fail in contexts like these?

When someone needs to find an address quickly, Google Maps is invariably the first application they open. In non-crisis times, this often works reasonably well. People share GPS coordinates when available or describe locations relative to prominent landmarks Google recognizes. OLA Maps has emerged as a competitor, particularly popular among Indian users. Open-source initiatives conduct periodic mapping drives where volunteers systematically document house numbers, street names, and other polygons to populate OpenStreetMap. These crowdsourced efforts have genuinely improved coverage in some formal areas.

Yet through my volunteering work, I witnessed how these tools fail catastrophically during disasters. Google Maps exhibits 200+ meter errors on unstructured Indian addresses, far beyond the 50-100 meter precision emergency dispatch requires (6). More critically, existing geocoding services assume users can provide addresses in formats the services recognize. When someone tells me "the apartment near the flooded underbridge in Thanisandra," every word carries local meaning but none translates to computational coordinates. Traditional NLP cannot distinguish whether "behind temple" means north or south, nor which temple among hundreds sharing similar names. GPS coordinates require smartphone access and network connectivity, precisely what disasters eliminate first. The challenge multiplies when people communicate in other languages such as Tamil or Kannada, describing locations using local landmarks that existing tools cannot interpret for remote volunteers who lack both geographic and linguistic context. I watched volunteers resort to typing nearby landmarks into Google Maps, then verbally confirming over flaky phone lines whether they had identified the correct location. The process remained manual, time-consuming, and error-prone.

Traditional geocoding services are fundamentally limited by their design. They work by pattern matching against databases of known addresses and coordinates. When addresses don't follow recognizable patterns, when they are built on contextual relationships rather than structured formats,

these services have no mechanism to adapt. You cannot simply update Google Maps to understand "behind the old temple" without fundamentally changing how the technology processes language and context. A different approach is needed, one that can reason about spatial relationships the way humans do.

Large Language Models might bridge this divide. Unlike traditional NLP that extracts discrete entities, LLMs leverage contextual reasoning, resolve ambiguities through inference, interpret landmark relationships. Yet their crisis geocoding application remains entirely unexplored, particularly for low-resource, multilingual contexts where the need is most acute. Viability for resource constrained environments raises critical questions about infrastructure dependency, cost sustainability, and deployment feasibility. LLMs demand cloud connectivity and per-query API costs, potentially prohibitive when disasters compromise internet infrastructure (9) and budgets constrain local organizations. Most LLMs train predominantly on English data, reproducing linguistic biases disadvantaging non-English communications (10). Before advocating LLM solutions, we must evaluate not just accuracy but deployability, cost, and cultural appropriateness for organizations operating in precisely the contexts commercial services neglect.

Despite extensive research on crisis informatics (11) and geocoding challenges in developing contexts (5), existing work focusses primarily on formal address parsing or named entity recognition, neglecting the contextual reasoning capabilities LLMs might provide for unstructured crisis addresses. This research addresses that gap through empirical evaluation of LLM-augmented geocoding in resource-constrained, multilingual disaster response contexts. Specifically, I address three interconnected questions: Do current geocoding services work well enough? Can LLMs improve them? And if so, how do we deploy LLMs responsibly in humanitarian contexts?

We investigate through three complementary empirical studies using authentic crisis data:

RQ1: Baseline Geocoding Service Performance. *How accurately do contemporary geocoding services process unstructured crisis location descriptions, and do systematic performance differences emerge across operationally-relevant distance thresholds?*

We benchmark five major geocoding services (Google Maps, OLA Maps, OpenCage Geocoder, Nominatim, and Pelias) using 117 authentic flood crisis addresses from Bangalore, India, provided by the local voluntary organization coordinating relief operations. We evaluate performance across seven distance thresholds (50m to 5km) corresponding to operational requirements from building-specific rescue to general area dispatch.

RQ2: LLM-Based Address Augmentation Effectiveness. *Can Large Language Model-based address preprocessing improve geocoding accuracy for unstructured disaster locations, and which prompting strategies prove most effective?*

We evaluate eight augmentation techniques (zero-shot prompting, few-shot learning, retrieval-augmented generation, chain-of-thought reasoning, combined prompting, iterative refinement, role-based prompting, and deterministic processing) across the same 117 addresses.

We measure within-service improvement and cross-service ranking changes to identify optimal configurations for crisis deployment.

However, technical performance alone cannot determine whether this technology should be deployed. While technical performance metrics tell us whether LLM-augmented geocoding *works*, they don't reveal *who* it works for or what happens when we deploy these systems in actual humanitarian contexts. Crisis mapping platforms operate within complex power structures involving affected communities, humanitarian organizations, platform operators, and formal institutions. Automation changes these relationships in ways that efficiency metrics alone cannot capture.

Communities gain the ability to report in natural language but lose control over how their descriptions get interpreted. NGOs achieve operational efficiency but become dependent on AI infrastructure controlled by external entities. Platforms solve volunteer burnout but trade human adaptability for algorithmic consistency. These shifts raise questions that technical evaluation cannot answer: Who holds decision-making power in automated systems? How can communities challenge systematic errors? What governance mechanisms ensure accountability?

This leads to our third research question:

RQ3: What governance frameworks and deployment practices are necessary to ensure LLM-augmented geocoding serves humanitarian principles rather than simply optimizing existing power structures?

This question acknowledges that technology deployment is never neutral. It requires examining how benefits and risks distribute across stakeholders, what institutional arrangements enable genuine accountability, and whether automation reinforces or challenges existing inequalities in crisis response.

Detailed methodology for each research question appears in Chapter3. Results for RQ1 appear in Section4.1, RQ2 in Section4.2, and RQ3 in Section6.

Research Context and Constraints

This research has emerged from my five years of volunteer work with Robin Hood Army coordinating disaster relief, culminating in formal collaboration with the voluntary organization I worked with. The organization provided informed consent for using anonymized address data from their 2019 operations. All location data was de-identified; no personally identifiable information appears in our dataset. This research aims to support, not replace, the local knowledge and volunteer efforts I was part of.

Deployment realities constrain solution viability. During crises, internet connectivity fails, limiting cloud-dependent solutions. The organization operates on minimal budget with no technical staff, requiring low-cost, minimal-configuration solutions. Communications mix English, Kannada,

Tamil, and Hindi, but our LLM augmentation operates English-only (reflecting LLM availability and our linguistic constraints). Our knowledge base derives from OpenStreetMap data, enabling reproducibility but inheriting OSM's coverage biases toward formal areas over informal settlements (7).

These constraints shape evaluation criteria. We prioritize accuracy improvements alongside deployment feasibility, cost sustainability, and workflow integration. Superior accuracy proves useless if unavailable to local organizations during actual crises.

Contributions and Limitations

This research targets the geographic knowledge gaps that disadvantage informal settlements during disasters. Improved geocoding renders community-generated intelligence legible to formal systems by building computational bridges that preserve contextual richness without replacing local knowledge. Communities get faster response. Responders shrink search radii from kilometres to a few hundred meters. Remote volunteers partially compensate for missing local context, though technology cannot replace ground-truth verification.

Significant limitations constrain this work. English-only processing excludes most Indian crisis communications (Hindi, Tamil, Kannada, other languages). OpenStreetMap-derived knowledge bases inherit documented coverage gaps in informal settlements (7), precisely where challenges peak. Single-city (Bangalore), single-disaster (flooding) evaluation limits generalizability across India's diversity. Most critically, we cannot address root causes of addressing informality: land tenure insecurity, urban planning failures, state projects marginalizing populations (8). Technology can reduce symptoms. It cannot resolve structural inequalities.

This thesis contributes one component (improved geocoding) while recognizing technical solutions alone cannot address structural inequalities and resource constraints shaping disaster vulnerability in the global south.

2

Background and Related Work

Digital technologies have fundamentally reimagined how communities respond to disasters. No longer do we rely solely on hierarchical command structures and official channels. Instead, affected populations themselves have become vital sensors, reporters, and decision-makers in crisis situations. This remarkable shift toward participatory frameworks represents both a technological and philosophical transformation in understanding whose knowledge matters when catastrophe strikes.

2.1 Crisis Mapping and Location Challenges

Traditional cartography, once the exclusive domain of governments and corporations, has given way to something far more dynamic: crisis mapping powered by ordinary citizens (12). When Ushahidi emerged from Kenya's 2008 post-election violence, it represented a radical reimagining of how information flows during chaos. SMS messages, emails, and web reports converged into living maps that provided real-time updates from the ground (2, 12).

What makes contemporary crisis mapping revolutionary? Speed, for one. Long before humanitarian organizations can mobilize their machinery, local residents are already documenting unfolding disasters with reports of damage, trapped families, and urgent medical needs (2, 13). The old information vacuum that plagued early disaster response has been replaced by torrents of data. Citizens transform into sensors, their proximity to events making them invaluable witnesses (13, 14). Yet this same democratization breeds complexity.

The Haiti earthquake laid bare uncomfortable truths about crowdsourced crisis data. Of 40,000+ messages flooding the Ushahidi platform, volunteers could verify barely 3,500, a sobering statistic that speaks volumes about verification challenges (2). Worse still, categorization errors plagued 36% of messages, turning potentially life-saving information into misleading noise (2). These aren't merely technical glitches; they reflect deeper tensions between inclusive information gathering and the desperate need for accuracy when lives hang in balance.

Geographic precision remains persistently elusive. Consider how people actually describe locations during disasters: "near where the blue church used to stand" or "past the collapsed market, by

the big tree." Such descriptions, while perfectly clear to locals, remain computationally intractable for external geocoding systems (2, 14). This creates a problematic tradeoff: in pursuit of pinpoint accuracy, crisis mappers sometimes discard messages lacking precise coordinates, potentially losing critical information (2). External moderators, tasked with interpreting these colloquial references, inevitably impose their own spatial understanding, sometimes at odds with local knowledge systems (14).

This geocoding challenge poses significant barriers for crisis response effectiveness. When volunteers cannot reliably convert informal location descriptions into coordinates, response time increases and aid may be misdirected. Understanding how different geocoding services handle unstructured addresses becomes critical for evaluating crisis mapping platform capabilities. Yet systematic comparisons of geocoding service performance on crisis-context addresses remain limited in existing literature, particularly for South Asian contexts where informal addressing systems dominate. Moreover, the potential for automated augmentation techniques to standardize informal addresses before geocoding has received little empirical evaluation, despite advances in natural language processing capabilities.

Perhaps most troubling is the phenomenon of "dead maps," platforms launched with enthusiasm only to wither from neglect. Among nearly 13,000 Ushahidi deployments analyzed, countless showed minimal activity or remained entirely uncustomized (15). Without community credibility and local ownership, even the most sophisticated platforms become irrelevant (14, 15). This sustainability challenge reveals deeper ICT4D concerns about technology adoption in resource-constrained contexts. Platforms fail not due to technical inadequacy but from misalignment with local governance structures, insufficient training for community moderators, and lack of ongoing technical support. Deployment feasibility extends beyond initial setup to include long-term maintenance costs, infrastructure dependencies, and organizational capacity to sustain operations. These considerations become particularly acute when evaluating newer computational approaches that may demand cloud connectivity, ongoing API expenses, or specialized technical expertise unavailable to grassroots organizations.

2.2 Geocoding Services in Disaster Contexts

Geocoding during disasters operates under extraordinary constraints that would challenge any system, particularly when processing frantic, unstructured messages from traumatized populations (2, 14). Infrastructure collapses. Reference points vanish. The very landmarks that anchor spatial understanding (buildings, bridges, familiar streets) may no longer exist.

Haiti's experience illuminated these challenges starkly. Location references required more than simple translation; they demanded cultural and spatial interpretation that remote volunteers, however dedicated, struggled to provide (2, 16). A global collaboration emerged, connecting translators, geocoders, and mapping experts across continents, yet this distributed approach revealed fundamental limitations in processing locally-specific spatial knowledge (16).

Technical infrastructure buckles under crisis conditions. The Czech Republic's 2013 floods saw servers fail catastrophically from overload, forcing rapid adjustments to caching and processing systems while flood waters rose (13). Such failures represent critical vulnerabilities that emerge precisely when spatial accuracy matters most.

A deeper challenge lies in the representational logic of crisis mapping itself. These platforms require precise Cartesian coordinates, yet communities often conceptualize space through relationships, memories, and informal landmarks (14). "Two houses past the temple" makes perfect sense locally but confounds standardized geocoding. This represents not merely a technical problem but an epistemological clash between different ways of knowing and describing space.

Mobile technology promises solutions through GPS-enabled devices that automatically capture locations (14, 17). But this technological fix creates new inequalities. Those with simpler phones, or in areas where GPS signals struggle, become invisible to the crisis map. The geographic representation skews toward technologically privileged populations, potentially misdirecting aid away from the most vulnerable (14, 17).

2.3 Large Language Models for Spatial Understanding

Existing crowdsourcing platforms have not yet fully embraced Large Language Models, though current approaches already demonstrate sophisticated natural language processing capabilities (18, 19, 20). Ushahidi's SwiftRiver, for instance, employs NLP and AI techniques to filter and structure the chaotic streams of crisis information, suggesting readiness for more advanced language understanding (18, 19, 20).

Geographic Information Retrieval techniques currently dominate spatial information processing in these contexts. TF-IDF vector space models assess thematic relevance, while systems like Lucene parse crowdsourced content for disaster-relevant information (21, 22). These approaches work, but they essentially perform sophisticated pattern matching rather than true language understanding.

Semantic analysis capabilities are creeping into platforms, enabling auto-categorization that hints at deeper comprehension (20). Probabilistic models, borrowing techniques from spam detection, now assess message credibility and identify damage patterns at both microscopic and macroscopic scales (22). Each advancement edges closer to the kind of contextual understanding that LLMs could provide.

Yet significant hurdles remain. Local languages twist and evolve, carrying spatial references that no model trained on standard corpora would recognize (21, 22). Disaster response demands instantaneous processing, potentially eliminating computationally intensive models from consideration. The balance between accuracy and speed becomes critical for effective crisis response.

Sentiment analysis has found its way into disaster SMS processing. The European Council's Joint Research Center analysed Haiti's message streams not just for facts but for emotional temperature, recognizing that population sentiment itself constitutes critical intelligence (18, 21). This broader

understanding of communication (beyond mere information extraction) suggests pathways for LLM integration.

But realizing this potential requires confronting computational constraints, linguistic diversity, and the fundamental challenge of encoding local spatial knowledge into models trained on global data. The question isn't whether LLMs will transform crisis geocoding, but whether they can do so while respecting the nuanced, culturally-specific ways communities understand and describe their own spaces.

2.4 Summary and Research Gap

The literature reveals a persistent tension at the heart of crisis mapping: platforms designed to democratize disaster response often struggle with the very informality that makes crowdsourced intelligence valuable. Citizens report locations using landmarks, relationships, and colloquial references that defy standardized geocoding. Haiti demonstrated this challenge at scale, thousands of potentially life-saving messages remained unmapped because remote volunteers lacked contextual knowledge to interpret local spatial descriptions. The problem intensifies in South Asian contexts, where addressing systems evolved through landmark-based wayfinding rather than administrative grids, creating systematic barriers for populations already marginalized by infrastructure gaps.

Yet research addressing these geocoding failures remains surprisingly sparse. We lack systematic evaluations comparing how different geocoding services handle unstructured crisis addresses, particularly in developing contexts where informal addressing dominates. Do contemporary geocoding APIs differ meaningfully in their ability to parse colloquial location descriptions? Which services prove most robust when landmarks replace street numbers, when misspellings proliferate under stress, when administrative hierarchies collapse into concatenated strings? Without comparative benchmarks, crisis mapping platforms cannot make informed choices about geocoding infrastructure.

More critically, the potential for computational augmentation remains under explored. Large Language Models demonstrate contextual reasoning capabilities that could bridge the gap between informal descriptions and structured addresses. Can LLMs standardize unstructured location text before geocoding, improving accuracy where traditional parsing fails? Which prompting strategies prove effective, zero-shot instructions, few-shot learning, knowledge-augmented approaches? Existing crisis informatics research acknowledges NLP's role in filtering and categorizing crowdsourced data but stops short of evaluating whether modern language models can tackle the geocoding preprocessing challenge specifically.

Finally, technical capability means little without deployment viability. The "dead maps" phenomenon warns against assuming sophisticated technologies will translate into sustained adoption. Even if LLM augmentation improves geocoding accuracy, does it remain feasible for resource-constrained voluntary organizations operating under disaster conditions? What infrastructure

dependencies emerge, cloud connectivity, API costs, technical expertise requirements? Can grassroots responders actually integrate such approaches into existing workflows, or do computational demands render them impractical precisely when they are needed most?

This thesis addresses these interconnected gaps through three research questions. RQ1 establishes baseline performance by systematically comparing five geocoding services on authentic Bangalore flood response addresses, quantifying accuracy across operationally meaningful distance thresholds. RQ2 evaluates whether LLM-based preprocessing can improve geocoding outcomes, testing eight augmentation techniques to identify effective prompting strategies for crisis-context address standardization. RQ3 moves beyond accuracy to examine deployment feasibility, measuring computational costs, infrastructure requirements, and workflow integration challenges for voluntary organizations with limited technical capacity. Together, these questions shift focus from abstract technical potential toward grounded assessment of what actually works in resource-constrained disaster response settings.

3

Methodology

3.1 Dataset

The evaluation dataset consisted of 117 real-world unstructured location descriptions shared by Robin Hood Army, a voluntary organization, during flood response scenarios in Bangalore, India. This corpus represents authentic crisis communication data where location information is often imprecise, hurried, or relies on local landmarks rather than formal addressing conventions.

3.1.1 Dataset Characteristics

Geographic Scope:

The study focuses on the Bangalore metropolitan area (12.9716°N, 77.5946°E), selected for its diverse addressing challenges including:

- Mix of formal and informal settlements
- Multilingual street naming (English, Kannada)
- Rapid urban expansion with incomplete address infrastructure
- Coexistence of traditional landmarks and modern GPS-based navigation

Address Type Distribution

The corpus encompasses three categories of location descriptions typical in crisis contexts:

1. Formal Structured Addresses (e.g., "123 MG Road, Indiranagar, Bangalore 560038")
2. Informal Landmark-Based Descriptions (e.g., "near water tank behind the temple", "opposite BMTC bus depot")
3. Partial/Incomplete Addresses (e.g., "Koramangala 5th Block", "HSR Layout Sector 2")

3.1.2 Ground Truth Establishment

Each of the 117 locations was manually verified to establish ground truth coordinates through a multi-stage validation process:

1. **Satellite Imagery Analysis:** High-resolution imagery from Google Earth used to identify precise locations
2. **Local Knowledge Verification:** Locations validated by researchers familiar with Bangalore geography
3. **Uncertainty Documentation:** Cases with positional ambiguity flagged for conservative error tolerance

Ground Truth Precision

Manual verification is estimated to have positional accuracy within ±10-20 meters for most locations, with higher uncertainty (±50m) for vague landmark descriptions.

Address Completeness Classification

Each address was binary-classified as:

- **Complete**: Contains all major hierarchical components (street, area, city) necessary for unambiguous identification
- **Incomplete**: Missing critical components, relies on landmarks, or provides only area-level information

3.1.3 Dataset Representativeness

The dataset represents realistic crisis communication patterns:

- **Time pressure**: Reports submitted rapidly without address verification
- **Citizen reporting**: Descriptions from non-professional observers using familiar landmarks
- **Infrastructure gaps**: Many locations lack formal addresses or GPS awareness
- **Linguistic variation**: Mix of English and transliterated local language names

Limitations:

- Single geographic region (Bangalore); findings may not generalize to other addressing systems
- Sample size (n=117) sufficient for statistical significance but larger corpus would strengthen external validity
- Temporal snapshot; addressing infrastructure and service performance evolve over time

3.2 Geocoding Service Evaluation Framework

This section describes the comprehensive evaluation framework designed to assess geocoding service performance when processing unstructured location descriptions from disaster contexts.

3.2.1 Geocoding Services Selection

We selected five geocoding services representing diverse technical approaches and market segments:

Table 3.1: Selected Geocoding Services

Service	Type	Provider	Rationale
Google Maps Geocoding API	Commercial	Google	Industry standard; extensive global coverage
Nominatim	Open-source	OpenStreetMap Foundation	Most widely deployed open-source geocoder
Pelias	Open-source	Linux Foundation	Modern modular architecture
OLA Maps	Commercial	ANI Technologies (OLA)	India-specific service with regional optimizations
OpenCage Geocoder	Commercial	OpenCage GmbH	Aggregator combining multiple sources

Selection Criteria:

- **License diversity**: Commercial vs. open-source options for cost/sustainability analysis
- **Geographic focus**: Global services (Google, Nominatim) vs. regional specialist (OLA)
- **Technical approach**: Single-source (Google, OLA) vs. aggregated (OpenCage)
- **Deployment feasibility**: Services available via API or self-hosted infrastructure

3.2.2 Evaluation Metrics

The evaluation framework employs five metric categories to comprehensively assess geocoding performance:

3.2.2.1 Coverage and Availability

Service Availability Rate captures the percentage of input addresses yielding valid geocoding results:

$$\text{Coverage (\%)} = \frac{\text{Valid Results}}{\text{Total Addresses}} \times 100 \tag{3.1}$$

In crisis scenarios, failing to return any location is often worse than returning an imprecise one. Coverage thus represents the "first-stage filter" of service utility, high accuracy becomes irrelevant if the service cannot process unstructured inputs.

3.2.2.2 Positional Accuracy and Precision

We calculate distance errors using the **Haversine great-circle distance** between geocoded coordinates and ground truth:

$$d = 2r \times \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{3.2}$$

where φ represents latitude, λ represents longitude (both in radians), $r = 6,371$ km (Earth's radius), and $\Delta\varphi$ and $\Delta\lambda$ denote coordinate differences.

Central Tendency Metrics: We report median error (CEP50) as the primary metric due to its robustness against outliers, alongside mean error and geometric mean for distribution characterization.

Dispersion Metrics: RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) assess error spread, with RMSE emphasizing large deviations. Standard deviation, IQR (Interquartile Range), and coefficient of variation provide complementary views of consistency.

Percentile Analysis: CEP95 (95th percentile) establishes the threshold below which 95% of errors fall, while 90th and 99th percentiles capture worst-case performance, critical for crisis response planning.

Precision at Distance Thresholds tracks the percentage of results within operationally meaningful distances:

$$\text{Precision@Threshold} = \frac{\text{Count}(\text{error} \leq \text{threshold})}{\text{Valid Results}} \times 100 \tag{3.3}$$

Different crisis response actions demand varying positional precision. Resource dispatch may tolerate 500m uncertainty, while building-specific rescue operations require sub-100m accuracy.

3.2.2.3 Geographic Bias and Spatial Patterns

Directional Bias exposes systematic tendencies to geocode locations in particular cardinal directions:

- **North-South Bias**: mean($\text{latitude}_{\text{geocoded}}$ - $\text{latitude}_{\text{truth}}$), converted to meters
- **East-West Bias**: mean($\text{longitude}_{\text{geocoded}}$ - $\text{longitude}_{\text{truth}}$), converted to meters

Spatial Correlation examines whether error magnitude correlates with distance from city centre (12.9716°N, 77.5946°E):

$$r = \text{corr}(\text{error}, \text{distance}_{\text{from centre}}) \tag{3.4}$$

Strong positive correlation ($r > 0.3$) suggests accuracy degrades away from urban core. Weak correlation ($|r| < 0.2$) signals consistent performance across the region. Negative correlation ($r < -0.3$), though rare, implies better suburban performance. Geographic bias analysis reveals whether

services systematically favour certain areas, an equity concern in crisis contexts where marginal areas often face greater vulnerability.

3.2.2.4 Quality Assessment and Statistical Validation

Outlier Detection employs multiple methods:

1. **Z-score Method**: Outliers exceed 2 standard deviations ($|z| > 2$), extreme outliers exceed 3 ($|z| > 3$)

2. **IQR Method**: Outliers fall outside $[Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}]$

3. **Domain-Specific Threshold**: Extreme errors exceed 5 km (likely catastrophic failures)

Data Quality Score $= 100 -$ (Z-score outlier rate), though this metric can mislead when entire distributions shift (as with poorly performing services).

Statistical Significance Testing determines whether observed performance differences reflect genuine disparities or random variation:

- **Mann-Whitney U Test** (non-parametric): Compares whether two services' error distributions differ significantly ($\alpha = 0.05$)

- **Kruskal-Wallis H Test**: Omnibus test assessing whether at least one service differs from others

- **Cohen's d Effect Size**: Quantifies practical significance of differences ($|d| < 0.2$: negligible; $0.2 \leq |d| < 0.5$: small; $0.5 \leq |d| < 0.8$: medium; $|d| \geq 0.8$: large)

Statistical significance (p-value) signals a difference exists; effect size reveals whether that difference matters practically. We employ non-parametric tests given the expected non-normal, right-skewed error distributions.

3.2.2.5 Robustness to Input Quality

Stratified Analysis by Address Completeness calculates all metrics separately for complete versus incomplete addresses, revealing service tolerance for degraded inputs, crucial for crisis scenarios where formal addresses may be unavailable.

Robustness Score $= \frac{\text{Incomplete Performance}}{\text{Complete Performance}} \times 100$, with higher scores indicating maintained effectiveness despite input quality degradation.

3.2.3 Experimental Procedure

Stage 1: Ground Truth Validation

1. Manual coordinate verification for all 117 locations using satellite imagery
2. Cross-validation with local geographic knowledge
3. Documentation of positional uncertainty estimates
4. Address completeness classification

Stage 2: Geocoding Execution

1. Query each service with identical unstructured address strings
2. Record returned coordinates (latitude, longitude)
3. Document null responses, errors, or invalid results
4. Timestamp queries to control for temporal service variations

Stage 3: Error Calculation

1. Compute Haversine distance between service coordinates and ground truth
2. Generate distance error matrix: 117 locations $\times$ 5 services
3. Flag extreme errors (>5km) for manual inspection

Stage 4: Statistical Analysis

1. Calculate all accuracy, precision, and quality metrics per service
2. Perform pairwise statistical comparisons (Mann-Whitney, Cohen's d)
3. Execute Kruskal-Wallis test for multi-service comparison
4. Compute geographic bias and spatial correlation metrics
5. Conduct stratified analysis by address completeness

Stage 5: Visualization and Interpretation

1. Generate comparative plots (box plots, error distributions, precision curves)
2. Create geographic bias visualizations (error vectors, heat maps)
3. Produce statistical summary tables
4. Interpret findings in crisis response context

3.2.4 Evaluation Framework Validation

The framework design addresses key validity concerns. **Internal and construct validity** emerge through ground truth established independently of any geocoding service, identical queries sent to all services (eliminating input variation), and multiple complementary metrics linked directly to operational crisis response requirements. We selected thresholds drawing from disaster management literature and practitioner input, while jointly evaluating coverage and accuracy avoids optimizing one at the other's expense.

External and statistical validity stem from using real-world addresses from actual crisis deployments with representative address type distributions for disaster contexts. The geographic focus on Bangalore captures addressing challenges common to many developing urban areas. Non-parametric statistical tests prove appropriate for non-normal distributions, with effect sizes reported alongside significance tests to distinguish statistical from practical differences.

3.3 LLM-Based Address Augmentation Evaluation Framework

This section describes the evaluation framework for assessing whether Large Language Model-based address preprocessing enhances geocoding accuracy when processing unstructured crisis location descriptions. While Section 3.2 established baseline service performance, we now evaluate LLM augmentation effectiveness through within-service improvement analysis and cross-service performance comparison.

3.3.1 LLM Augmentation Techniques

3.3.1.1 Prompts

Eight augmentation techniques address common deficiencies in crisis-context addresses: missing landmarks, informal descriptions, spelling variations, and incomplete administrative hierarchies. We employ GPT-OSS-20B-1 via AWS Bedrock with structured output constraints ensuring parseable JSON responses.

Table 3.2: LLM Augmentation Techniques for Address Standardization

ID	Description
T1	**Zero-Shot Baseline**: Apply basic standardization steps (split concatenated tokens, correct spelling, identify building numbers, remove delivery instructions and phone numbers) while preserving core location components. Output format structured as building→street→sublocality→locality→city→state→PIN→country.

Continued on next page

Table 3.2 – *Continued from previous page*

ID	Description
T2	**Few-Shot Learning**: Demonstrate standardization patterns through three diverse example transformations showing proper handling of concatenated tokens, door numbers, and landmark-based descriptions. The model learns from examples before processing the target address.
T3	**Context-Augmented RAG**: Provide Bangalore-specific knowledge base containing locality hierarchies, landmark aliases, common misspellings, and regional terminology. The model applies this contextual knowledge to correct misspellings, expand aliases, identify address chunks, and reconstruct proper administrative hierarchy.
T4	**Combined System Prompting**: Assign expert role (address standardization specialist for Bangalore), provide knowledge base context, and specify explicit protocol: de-concatenate tokens, spell-check against knowledge base, identify chunks using terminating tokens, reconstruct hierarchy (building→street→sub-locality→locality→city→state), and add missing administrative components.
T5	**Chain-of-Thought Reasoning**: Request step-by-step reasoning through standardization process with brief explanations (1-2 sentences per step): tokenize and split concatenated tokens, correct misspellings, identify address chunks, remove extraneous information, and reconstruct in hierarchical order.
T6	**Iterative Refinement**: Initial pass uses T4 (Combined) technique. Subsequent iterations receive geocoding feedback (previous result, positional error distance) and apply refinements targeting error reduction. For errors exceeding 500m, prompts suggest adding specific components or alternative aliases.
T7	**Role-Based Expert Persona**: Invoke expert identity (Dr. Priya Sharma, GIS specialist with 15 years Bangalore experience) with specific expertise (addressing conventions, regional terminology, linguistic variations, urban planning layouts) and mission context (emergency response system requiring precise locations where lives depend on accuracy).
T8	**Deterministic Rule-Based Protocol**: Execute exact rule sequence at temperature=0 for maximum consistency: normalize (lowercase, remove special characters except , / : #, collapse spaces), process tokens (split concatenated, correct misspellings, expand abbreviations), extract components (building number, road, locality indicators), filter non-location information (phone numbers, delivery instructions), and assemble in standard format.

These eight techniques span diverse prompting strategies from recent LLM research. T1 establishes the zero-shot baseline with explicit instructions but no examples. T2 applies few-shot learning, demonstrating patterns through examples. T3 employs retrieval-augmented generation

(RAG) with domain-specific knowledge. T4 combines system prompting, knowledge context, and structured protocols. T5 leverages chain-of-thought reasoning with explicit step articulation. T6 implements iterative refinement using geocoding feedback. T7 applies role-based prompting with expert personas and mission framing. T8 uses deterministic rule-based processing at temperature=0 for maximum reproducibility.

All techniques constrain geographic scope to the Bangalore metropolitan region and mandate structured JSON output containing only the standardized address. Instructions explicitly prohibit information hallucination when geographic knowledge proves uncertain. We set temperature to 0.2 for T1-T7, balancing determinism with linguistic flexibility, while T8 operates at temperature=0 for strict reproducibility. This design variety enables evaluation across the full spectrum of LLM augmentation approaches, from minimal intervention (T1, T8) through knowledge-enhanced processing (T3, T4) to sophisticated reasoning strategies (T5, T6, T7).

Complete Prompt Templates: Full prompt text for all eight techniques appears in Appendix 9.1, enabling reproducibility and detailed inspection of prompting strategies. The appendix includes exact wording, few-shot examples (T2), knowledge base content (T3, T4), and structured instructions for each technique.

3.3.1.2 Knowledge Base Construction

Techniques T3 (Context-Augmented RAG) and T4 (Combined System Prompting) incorporate a Bangalore-specific knowledge base to provide domain context for address standardization. The knowledge base comprises five categories:

Terminating tokens identify boundaries between address components. Five categories capture regional addressing conventions: road identifiers (road, marg, cross, street), locality identifiers (nagar, layout, colony, sector, stage), landmark identifiers (temple, park, signal, junction), commercial identifiers (mall, complex, tower, building), and specific road names (Sarjapur Road, Bannerghatta Road, Hosur Road). These tokens enable chunking of concatenated address strings into hierarchical components.

Regional aliases map informal names to canonical forms: HSR/HSR Layout, BTM/BTM Layout, Kormangala/Koramangala. The knowledge base contains 110 alias mappings extracted from OpenStreetMap data, local community forums, and real estate listings. Aliases address phonetic variations, abbreviations, and colloquial references common in crisis communications.

Common misspellings correct frequent errors: Bangaluru/Bangalore, Yellahanka/Yelahanka, Indranagar/Indiranagar. Ten misspelling patterns identified through manual inspection of training data enable standardization of crisis reports typed under stress or transmitted through error-prone communication channels.

Area hierarchies define administrative relationships: Koramangala contains sectors 1-7, HSR Layout contains sectors 1-6 and BDA Complex area. Seven major area hierarchies guide reconstruction of partial addresses missing locality or sub-locality components.

Data sources and construction process: Knowledge base population proceeded through systematic data collection and curation:

OpenStreetMap extraction. We downloaded Bangalore administrative boundaries, landmark names, and road networks from OpenStreetMap using the Overpass API with geographic bounding box coordinates (12.7°N-13.2°N, 77.4°E-77.8°E) covering the Bangalore metropolitan region. The raw OSM XML export contained thousands of entities including roads, buildings, amenities, and administrative boundaries. We filtered relevant features using OSM tags: `highway=*` for roads, `place=*` for localities, `amenity=*` for landmarks, and `boundary=administrative` for area hierarchies.

Data cleaning and standardization. Raw OSM data required substantial cleaning. Road names appeared inconsistently ("Sarjapur Road" vs "Sarjapura Road" vs "Sarjapur Rd"), requiring canonicalization to single preferred forms. Locality names contained spelling variations, transliteration differences (Kannada to English), and formatting inconsistencies. We deduplicated entries, standardized capitalization, removed special characters that might confuse LLM parsing, and selected the most commonly used variant for each location based on OSM usage frequency metadata.

Manual curation and augmentation. OSM data alone proved insufficient for crisis address patterns. We manually added: (1) common misspellings identified through inspection of training data addresses (10 frequent patterns covering phonetic errors, dropped letters, and transliteration variations), (2) colloquial aliases from local community forums and real estate platforms (HSR for HSR Layout, BTM for BTM Layout, informal sector numbering conventions), (3) regional terminology from official Karnataka State addressing guidelines (proper usage of "nagar," "layout," "stage," and "sector" hierarchies), and (4) crisis-specific landmarks referenced in disaster communications but absent from OSM (temporary shelters, flood-prone areas, evacuation points).

Hierarchical organization. Area hierarchies required manual structuring since OSM's flat tagging system doesn't capture containment relationships explicitly. We defined parent-child relationships (e.g., Koramangala contains Koramangala 1st Block through 7th Block, HSR Layout contains Sector 1 through Sector 6 plus BDA Complex) by cross-referencing administrative boundaries with ground truth addresses from our crisis dataset. This hierarchy enables reconstruction of incomplete addresses missing intermediate administrative levels.

The final knowledge base stored in `bangalore_kb.json` contains approximately 300 entries across all categories: 50+ terminating token patterns, 110 regional alias mappings, 10 common misspelling corrections, 50+ landmark names, and 7 major area hierarchies with sub-components. This size balances comprehensiveness with compactness, fitting within GPT-OSS-20B-1's context window when included in T3 and T4 prompts without exceeding token limits.

3.3.2 Evaluation Framework Design

The augmentation evaluation employs a two-part analysis framework addressing distinct research questions:

Part 1: Within-Service Improvement Analysis. For each geocoding service (Google Maps, OLA Maps, OpenCage Geocoder), we compare baseline performance (original address input) against augmented performance (LLM-processed address input) across all eight techniques. This design isolates augmentation effectiveness from inherent service differences, answering: *Does LLM augmentation improve this service's accuracy?*

For each service-technique pair (s, t), comparison examines error change:

$$\Delta_{\text{error}}(s,t) = \text{Error}_{\text{baseline}}(s) - \text{Error}_{\text{augmented}}(s,t) \tag{3.5}$$

Positive values suggest augmentation reduces positional error; negative values signal degradation. Statistical tests determine whether observed differences reflect genuine improvement or random variation.

Part 2: Cross-Service Performance Comparison. At each augmentation technique level (including baseline), we compare absolute performance across all three services. This approach identifies service ranking changes introduced by augmentation, answering: *Which service performs best at this technique level?*

For each technique t (including t_0 = baseline), service rankings emerge from precision at operational thresholds:

$$\text{Rank}(s,t) = \text{rank}_{\text{descending}}\left(\text{Precision}_{@500\text{m}}(s,t)\right) \tag{3.6}$$

Rankings may shift across techniques, revealing whether augmentation can elevate mid-tier services above baseline high-performers.

This two-part framework provides comprehensive evaluation: within-service analysis quantifies improvement magnitude per service, while cross-service analysis reveals competitive positioning changes and identifies optimal service-technique configurations for deployment.

3.3.3 Evaluation Metrics

Six complementary analyses assess augmentation effectiveness across coverage, accuracy, precision, and statistical reliability dimensions.

3.3.3.1 Coverage Analysis

Coverage captures the percentage of addresses returning valid geocoding results. Augmentation that improves accuracy while reducing coverage proves counterproductive, since failing to return any location is often worse than returning an imprecise one in crisis contexts. We calculate coverage per service-technique pair:

$$\text{Coverage}(s,t) = \frac{|\{\text{valid results}\}|}{|\{\text{total addresses}\}|} \times 100\% \tag{3.7}$$

Change in coverage reveals whether augmentation introduces parsing failures:

$$\Delta\text{Coverage}(s,t) = \text{Coverage}_{\text{augmented}}(s,t) - \text{Coverage}_{\text{baseline}}(s) \tag{3.8}$$

Negative values signal that augmentation harms coverage, rendering techniques counterproductive regardless of accuracy improvements.

3.3.3.2 Positional Accuracy Summary Statistics

For addresses returning valid results under both baseline and augmented conditions, we compute summary statistics of positional error (Haversine distance from ground truth):

- **Median error** (CEP50): Robust central tendency measure, resistant to outliers
- **Mean error**: Expected error magnitude
- **Standard deviation**: Error variability indicating consistency
- **Minimum/maximum error**: Range bounds

Median improvement assesses augmentation effectiveness:

$$\text{Improvement}_{\text{median}}(s,t) = \frac{\text{Median}_{\text{baseline}}(s) - \text{Median}_{\text{augmented}}(s,t)}{\text{Median}_{\text{baseline}}(s)} \times 100\% \tag{3.9}$$

Positive percentages show augmentation reduces error; negative values demonstrate degradation.

3.3.3.3 Error Distribution Percentile Analysis

Percentile analysis exposes whether augmentation improves worst-case performance (critical for crisis response planning) or merely shifts central tendency while increasing variability:

- **CEP90** (90th percentile): 90% of results fall within this error bound
- **CEP95** (95th percentile): Upper tail performance
- **CEP99** (99th percentile): Catastrophic error threshold
- **IQR** (Interquartile range): $Q_3 - Q_1$, measures spread robustly

Techniques improving median but degrading CEP95/99 shift distributions favorably on average while paradoxically increasing worst-case errors unacceptable in operational deployment.

3.3.3.4 Precision at Distance Thresholds

Precision tracks the percentage of results within operationally meaningful distances. Seven thresholds span building-level to city-level accuracy: 50m, 100m, 200m, 500m, 1000m, 2000m, 5000m.

$$\text{Precision}_{@d}(s,t) = \frac{|\{\text{error}(s,t) \leq d\}|}{|\{\text{valid results}(s,t)\}|} \times 100\% \tag{3.10}$$

The 500m threshold represents the critical crisis response dispatch requirement. Precision change at this threshold quantifies operational impact:

$$\Delta\text{Precision}_{@500\text{m}}(s,t) = \text{Precision}_{\text{augmented}}(s,t) - \text{Precision}_{\text{baseline}}(s) \tag{3.11}$$

Positive values show augmentation increases the count of operationally acceptable results. Techniques yielding positive median improvement but negative precision change at 500m may reduce average error while paradoxically decreasing counts of deployment-ready locations.

3.3.3.5 Service Rankings by Threshold

For each technique t (including baseline), we rank services by precision at each threshold. Rankings capture competitive positioning:

$$\text{Rank}_{@d}(s,t) \in \{1,2,3\} \quad \text{where } 1 = \text{best} \tag{3.12}$$

Rank changes across techniques expose whether augmentation can elevate mid-tier services above baseline leaders. Cross-technique consistency in rankings suggests robust service superiority independent of preprocessing, while rank volatility indicates augmentation-dependent performance requiring careful technique selection.

3.3.3.6 Statistical Significance Testing

Mann-Whitney U tests compare baseline versus augmented error distributions for each service-technique pair ($\alpha = 0.05$). This unpaired non-parametric test requires no distributional assumptions, proving appropriate given anticipated skewed error distributions with outliers. We employ the unpaired formulation to accommodate addresses with valid results under only one condition (baseline or augmented), maximizing sample utilization when coverage differs between configurations.

Cohen's d effect size quantifies practical magnitude:

$$d = \frac{\bar{x}_{\text{baseline}} - \bar{x}_{\text{augmented}}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \tag{3.13}$$

Interpretation thresholds: $|d| < 0.2$ negligible, $0.2 \leq |d| < 0.5$ small, $0.5 \leq |d| < 0.8$ medium, $|d| \geq 0.8$ large.

Hypotheses. H_0 states LLM augmentation does not reduce positional error: $\text{median}_{\text{augmented}} \geq \text{median}_{\text{baseline}}$. H_A states augmentation reduces error: $\text{median}_{\text{augmented}} < \text{median}_{\text{baseline}}$. Rejecting H_0 ($p < 0.05$) combined with medium-to-large effect size constitutes strong evidence for augmentation effectiveness. Large percentage improvements lacking statistical significance suggest high variance or insufficient sample size.

Statistical significance and practical significance may diverge in small-sample crisis contexts. Techniques showing large practical improvements (e.g., 40% median error reduction) without statistical significance ($p > 0.05$) warrant consideration despite failure to meet traditional hypothesis testing thresholds, as Type II errors (false negatives) prove likely given limited datasets. Conversely, statistically significant improvements with negligible effect sizes ($d < 0.2$) may lack operational relevance despite p-value criteria satisfaction.

3.3.4 Experimental Procedure

Stage 1 - Address Augmentation. We apply each of eight augmentation techniques to all 117 addresses (complete and incomplete) using GPT-OSS-20B-1 via AWS Bedrock API. Recording augmented address text, modifications applied, and processing metadata allows quality validation. We validate outputs for format compliance (parseable JSON, required fields present) and document failures (API errors, malformed responses, timeouts). Total augmented addresses: $117 \times 8 = 936$.

Stage 2 - Geocoding with Augmented Addresses. We submit all 936 augmented addresses to the same three services evaluated in RQ1: Google Maps, OLA Maps, OpenCage Geocoder. Recording returned coordinates, null responses, and error conditions enables comprehensive coverage analysis. Maintaining temporal consistency through controlled query timing and consistent API versions ensures fair comparison. We implement rate limiting to respect service quotas.

Stage 3 - Error Distance Calculation. We calculate Haversine distance (meters) between ground truth coordinates and geocoded coordinates for both baseline and augmented results. This generates an error matrix: 117 addresses $\times$ 3 services $\times$ 9 configurations (1 baseline + 8 techniques) = 3,159 potential measurements (actual count lower due to geocoding failures).

Stage 4 - Part 1 Analysis. For each service-technique pair, we filter for addresses with valid results under at least one condition. Computing all metrics from Section 3.3.3 yields: coverage, summary statistics, percentiles, precision at thresholds, and statistical tests. We calculate improvement percentages and deltas. Total service-technique pairs analyzed: $3 \times 8 = 24$.

Stage 5 - Part 2 Analysis. For each technique level (including baseline), we compare absolute performance across all three services. Ranking services by precision at each threshold generates master comparison tables showing service performance at all configuration levels. We identify best service per technique and best technique per service. Total configurations analyzed: $1 + 8 = 9$.

3.3.5 Evaluation Framework Validity

Internal validity emerges through controlled augmentation procedure (same LLM model, same prompts, same temperature setting for all addresses) and multiple independent metrics preventing single-metric optimization artifacts. Comparing identical addresses across baseline and augmented conditions eliminates confounding factors while allowing different coverage rates under each configuration.

Construct validity stems from augmentation techniques grounded in geocoding best practices and crisis informatics literature, multiple complementary metrics capturing operational requirements (coverage, accuracy, precision, reliability), and explicit separation of statistical significance from practical effect size magnitude. We selected metrics based on crisis response deployment requirements rather than statistical convenience.

External validity considerations include geographic focus on Bangalore-specific contexts. Augmentation effectiveness may vary in other linguistic contexts (non-English, transliteration challenges), administrative systems (different address hierarchies), or geocoding service training data

distributions. The evaluation employs real crisis addresses from authentic flood response scenarios, representing genuine deployment challenges. Findings generalize to similar urban contexts in South Asia with English-language addresses but require validation before deployment in substantially different settings (rural contexts, non-English addressing systems, regions with different administrative hierarchies).

Threats to validity include LLM augmentation introducing non-determinism despite low temperature settings. Multiple augmentation runs on identical addresses may yield variations. We mitigate this through single-execution design reflecting realistic deployment, but acknowledge variability as inherent limitation. Temperature 0.2 balances determinism with linguistic flexibility requirements. Additionally, GPT-OSS-20B-1 training data recency may favor certain services' conventions, potentially biasing augmentation toward formats matching specific geocoder expectations. Cross-validation with alternative LLMs would strengthen generalizability but remains beyond current scope. The sample size (117 addresses) limits statistical power for detecting small effects, increasing Type II error risk. Power analysis suggests detecting medium effects requires this sample size, but small effects may evade detection despite practical relevance.

4

Results and Evaluation

4.1 Comparative Geocoding Performance

This section presents comprehensive evaluation results for five geocoding services processing 117 unstructured location descriptions from Bangalore, India crisis scenarios.

4.1.1 Service Coverage Analysis

Coverage rates capture each service's ability to return geocoding results for unstructured crisis addresses:

Table 4.1: Service Coverage and Availability

Service	Valid Results	Coverage (%)	Null Responses	Null Rate (%)
Google Maps	107	**91.5**	10	8.5
Pelias	107	**91.5**	10	8.5
OLA Maps	106	**90.6**	11	9.4
OpenCage	88	**75.2**	29	24.8
Nominatim	2	**1.7**	115	98.3

Google, Pelias, and OLA achieve coverage above 90%, successfully geocoding virtually all unstructured addresses. OpenCage returns valid results for 75.2% of addresses (88 of 117). Nominatim produces only 2 valid results from 117 addresses (1.7% coverage).

Methodological Note: Given Nominatim's coverage of 1.7% ($n = 2$), **subsequent comparative analysis focuses on the four services with adequate sample sizes**: Google Maps, OLA Maps, Pelias, and OpenCage. While Nominatim's two successful results show 97.9m median error, statistical analysis requires larger samples. We therefore exclude Nominatim from ranking discussions, statistical comparisons, and comparative findings throughout this section. For data completeness, Nominatim appears in tables with an asterisk noting insufficient sample size.

4.1.2 Positional Accuracy: Summary Statistics

Table 4.2 presents comprehensive accuracy metrics across services with valid results:

Table 4.2: Positional Accuracy Metrics (all values in meters except coverage)

Service	N	Cov. (%)	Median	Mean	Std Dev	RMSE	MAE	Geo. Mean	Min	Max
Google	107	91.5	**228.7**	1,033	2,173	2,397	1,033	161.8	0.4	16,227
OLA	106	90.6	**280.3**	1,180	2,978	3,190	1,180	194.1	3.9	23,812
Pelias	107	91.5	**7,181.7**	1,237,248	3,794,875	3,974,578	1,237,248	11,799	10.2	14,557,408
OpenCage	88	75.2	**9,420.4**	2,865,241	5,177,697	5,891,817	2,865,241	44,943	27.5	15,999,703
Nominatim*	2	1.7	**97.9**	97.9	99.7	120.6	97.9	68.0	27.5	168.4

*Excluded from analysis due to insufficient sample size ($n = 2$)

Google and OLA display median errors of 229m and 280m respectively. Minimum errors approach zero (Google: 0.4m, OLA: 3.9m). Maximum errors reach 16,227m (Google) and 23,812m (OLA).

Pelias and OpenCage exhibit median errors of 7,182m and 9,420m respectively. Maximum errors exceed 14,000,000m for both services.

Distribution characteristics: Geometric means fall substantially below arithmetic means for all services (Google: 162m geometric vs. 1,033m mean; Pelias: 11,799m geometric vs. 1,237,248m mean).

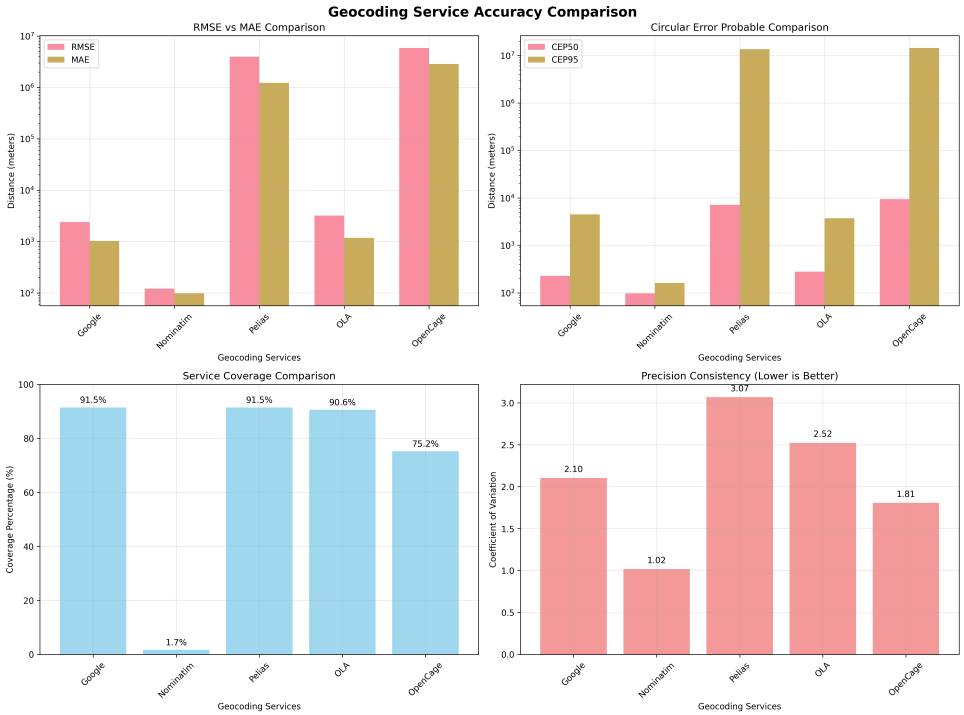

Figure 4.1: Comprehensive accuracy comparison showing distribution characteristics including median, quartiles, and outliers. Google and OLA: median errors ∼230-280m. Pelias and OpenCage: median errors exceeding 7km with numerous extreme outliers.

4.1.3 Error Distribution: Percentile Analysis

Percentile-based metrics quantify error bounds at different confidence levels:

Table 4.3: Percentile-Based Error Thresholds (meters)

Service	CEP50 (Median)	CEP95	90th %ile	99th %ile	IQR
Google	228.7	4,518.8	2,495.8	9,857.9	1,030.6
OLA	280.3	3,748.8	2,311.3	12,104.2	1,055.7
Pelias	7,181.7	13,508,718	1,898,340	14,529,195	11,344.7
OpenCage	9,420.4	14,463,641	14,039,565	15,929,859	3,094,294
Nominatim*	97.9	161.3	154.3	167.0	70.5

*Excluded from analysis due to insufficient sample size ($n = 2$)

Google and OLA maintain CEP95 values under 5 km (4,519m and 3,749m respectively), indicating 95% of geocoded locations fall within these error bounds. IQR values cluster around 1km for both services (1,031m and 1,056m).

Pelias and OpenCage display CEP95 values exceeding 13,000 km. Percentile values at 90th, 95th, and 99th levels remain in millions of meters for both services.

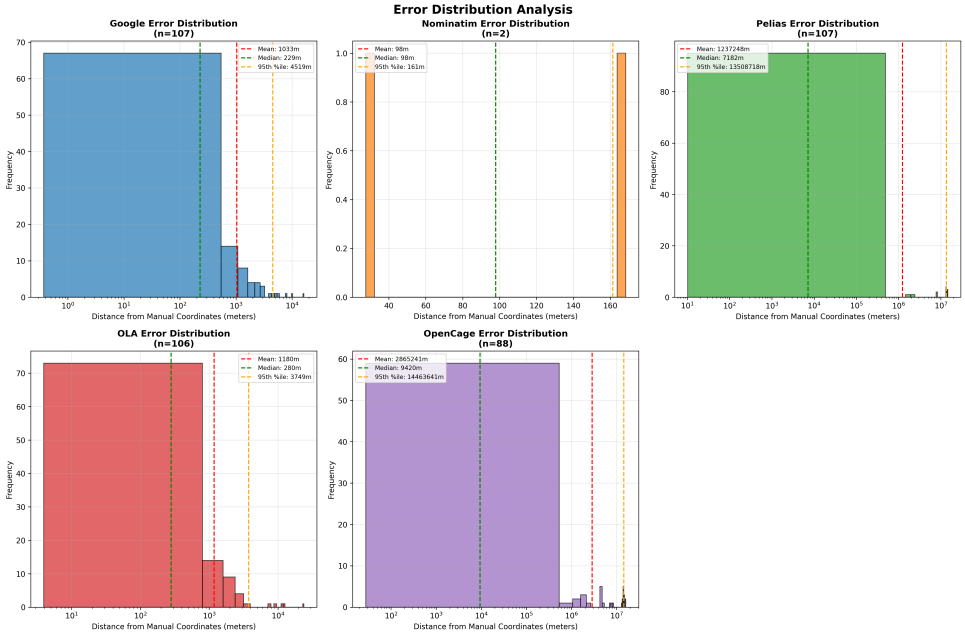

Figure 4.2: Distribution of positional errors across geocoding services. Google and OLA show right-skewed distributions with median errors <300m. Pelias and OpenCage exhibit bimodal distributions with majority of errors exceeding 7km. Box plots overlay median, quartiles, and outliers.

4.1.4 Precision at Distance Thresholds

Precision performance reveals the percentage of results falling within operationally meaningful distance thresholds:

Table 4.4: Precision Performance - Percentage of Valid Results Within Distance Threshold

Service	50m	100m	200m	500m	1000m	2000m	5000m
Google Maps	15.9%	**36.4%**	49.5%	**62.6%**	72.9%	86.9%	95.3%
OLA Maps	14.2%	**37.7%**	44.3%	**61.3%**	72.6%	87.7%	95.3%
OpenCage	0.0%	**2.3%**	3.4%	**8.0%**	12.5%	21.6%	37.5%
Pelias	0.0%	**0.9%**	1.9%	**5.6%**	10.3%	15.9%	34.6%
Nominatim*	0.0%	**50.0%**	**100.0%**	**100.0%**	100.0%	100.0%	100.0%

*Excluded from analysis due to insufficient sample size ($n = 2$)

At 500m threshold: Google achieves 62.6% precision, OLA reaches 61.3%, OpenCage attains 8.0%, Pelias achieves 5.6%.

At 50m threshold: Google reaches 15.9%, OLA achieves 14.2%, both OpenCage and Pelias reach 0%.

At 100m threshold: OLA leads with 37.7%, Google follows with 36.4%, OpenCage reaches 2.3%, Pelias attains 0.9%.

Precision progression for Google: 15.9% @ 50m → 36.4% @ 100m → 62.6% @ 500m → 72.9% @ 1km → 86.9% @ 2km → 95.3% @ 5km.

Precision progression for OLA: 14.2% @ 50m → 37.7% @ 100m → 61.3% @ 500m → 72.6% @ 1km → 87.7% @ 2km → 95.3% @ 5km.

Performance Summary by Threshold:

Table 4.5: Performance Summary by Threshold

Threshold	Best Service	Best Precision
50m	Google	**15.9%**
100m	OLA	**37.7%**
500m	Google	**62.6%**

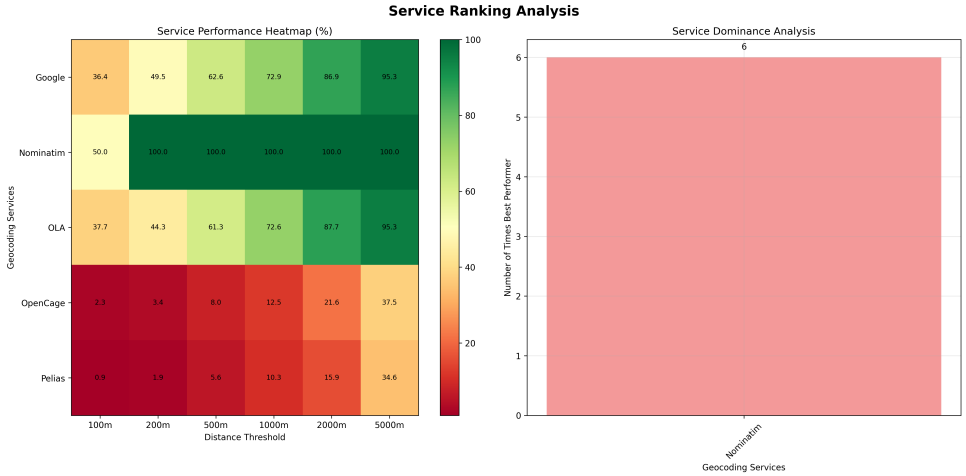

Figure 4.3: Precision performance curves showing percentage of results within distance thresholds. Google and OLA maintain superior performance across all thresholds, achieving ∼63% precision at 500m. Tier separation remains consistent from 50m through 5km thresholds.

4.1.5 Service Rankings by Threshold

Services ranked by precision at each threshold (1=best performance, Nominatim excluded due to $n = 2$):

Table 4.6: Service Rankings Across Distance Thresholds

Threshold	Rank 1	Rank 2	Rank 3	Rank 4
50m	Google (15.9%)	OLA (14.2%)	OpenCage (0%)	Pelias (0%)
100m	OLA (37.7%)	Google (36.4%)	OpenCage (2.3%)	Pelias (0.9%)
200m	Google (49.5%)	OLA (44.3%)	OpenCage (3.4%)	Pelias (1.9%)
500m	**Google (62.6%)**	**OLA (61.3%)**	OpenCage (8.0%)	Pelias (5.6%)
1000m	Google (72.9%)	OLA (72.6%)	OpenCage (12.5%)	Pelias (10.3%)
2000m	OLA (87.7%)	Google (86.9%)	OpenCage (21.6%)	Pelias (15.9%)
5000m	Google/OLA (95.3%)	Google/OLA (95.3%)	OpenCage (37.5%)	Pelias (34.6%)

Google and OLA exchange rank 1-2 positions across thresholds with differences typically under 5 percentage points. Tier separation between these leaders (60-95% precision) and Pelias/OpenCage (5-37% precision) remains consistent across all thresholds. At 5km threshold, Google and OLA maintain above 95% precision while Pelias and OpenCage achieve under 40%.

4.1.6 Statistical Significance Testing

Pairwise comparisons assess whether observed performance differences exceed random variation:

Table 4.7: Pairwise Statistical Test Results

Comparison	Mann-Whitney p	Significant (α=0.05)?	Cohen's d	Effect Size	Interpretation
Google vs OLA	0.830	**No**	-0.057	Negligible	No significant difference
Google vs Pelias	<0.001	**Yes**	-0.461	Small	Google significantly better
Google vs OpenCage	<0.001	**Yes**	-0.824	**Large**	Google significantly better
OLA vs Pelias	<0.001	**Yes**	0.460	Small	OLA significantly better
OLA vs OpenCage	<0.001	**Yes**	-0.822	**Large**	OLA significantly better
Pelias vs OpenCage	0.157	No	-0.364	Small	No significant difference

Google and OLA show no significant difference ($p = 0.830$, Cohen's $d = -0.057$ negligible effect). Both services significantly outperform Pelias and OpenCage ($p < 0.001$ for all comparisons). Effect sizes for Google/OLA versus OpenCage comparisons reach $d \sim 0.82$ (large). Pelias and OpenCage show no significant difference from each other ($p = 0.157$).

Multi-Service Comparison (Kruskal-Wallis H Test): $H = 45.23$, $p < 0.001$, confirming significant differences exist among services. Normality tests (Shapiro-Wilk $p < 0.05$ for all services) confirm non-normal, right-skewed error distributions.

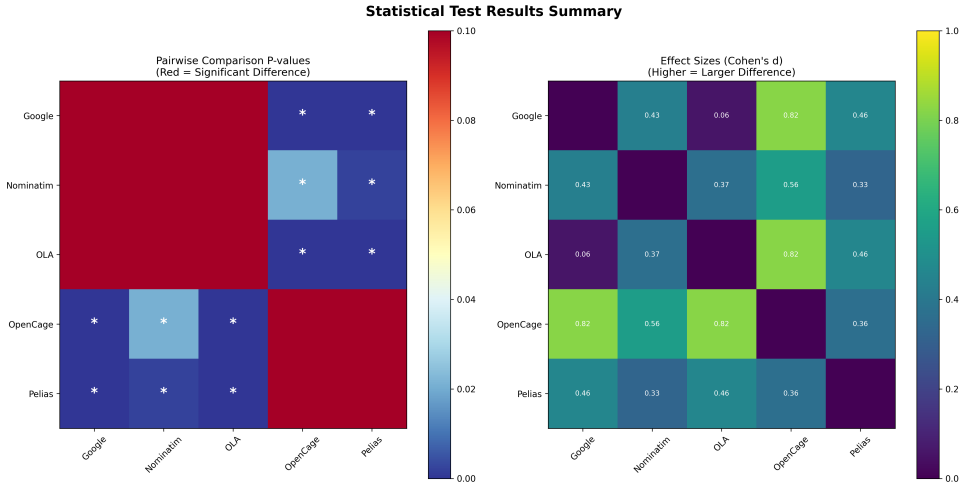

Figure 4.4: Statistical significance and effect size summary for pairwise service comparisons. Google and OLA show no significant difference ($p = 0.830$, $d = -0.057$). Both significantly outperform Pelias and OpenCage with large effect sizes ($d \sim 0.82$).

4.1.7 Geographic Bias and Spatial Patterns

Analysis of systematic directional biases and spatial accuracy patterns:

Table 4.8: Geographic Bias Metrics

Service	N-S Bias (m)	E-W Bias (m)	|Bias Vector| (m)	Dist-Center Corr.
Google	+38.9 (N)	+12.4 (E)	40.8	0.177
OLA	-363.3 (S)	+66.1 (E)	369.2	0.158
Pelias	+316,331 (N)	-1,507,245 (W)	1,540,097	0.108
OpenCage	+700,532 (N)	-3,324,074 (W)	3,397,169	-0.153
Nominatim*	+82.7 (N)	+9.4 (E)	83.2	1.000*

*Excluded from analysis due to insufficient sample size ($n = 2$)

Google exhibits bias vector magnitude of 40.8m with slight northward (+38.9m) and eastward (+12.4m) tendencies. OLA shows bias vector magnitude of 369.2m with moderate southward bias (-363.3m) and eastward bias (+66.1m).

Pelias displays bias vector magnitude of 1,540,097m (approximately 1,540 km) with strong northward (+316,331m) and westward (-1,507,245m) components. OpenCage exhibits bias vector magnitude of 3,397,169m (approximately 3,397 km) with strong northward (+700,532m) and westward (-3,324,074m) components.

Spatial correlation analysis: All services show weak correlations between error magnitude and distance from city center ($|r| < 0.2$ for Google, OLA, Pelias; $r = -0.153$ for OpenCage).

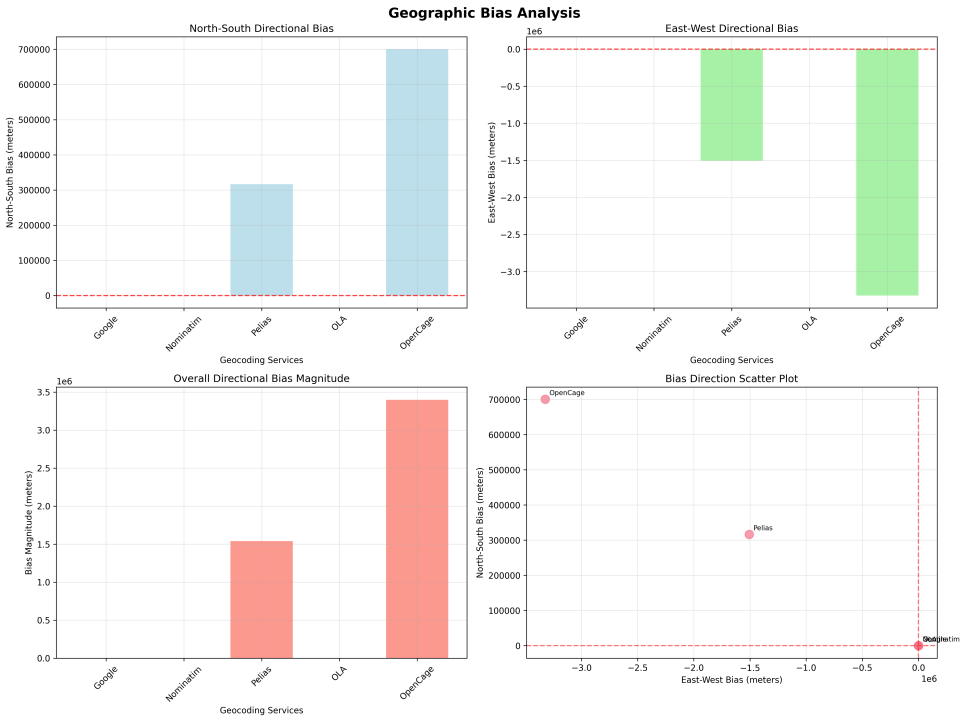

Figure 4.5: Geographic bias analysis revealing directional error patterns. Google: 40.8m vector magnitude. OLA: 369.2m southward bias. Pelias and OpenCage: biases exceeding 1,500km. Spatial correlation analysis reveals weak centre-periphery effects ($|r| < 0.2$ for most services).

4.1.8 Data Quality and Outlier Analysis

Assessment of result consistency and extreme error frequency:

Table 4.9: Outlier Rates and Quality Metrics

Service	Valid (N)	$>2\sigma$ (%)	$>3\sigma$ (%)	IQR (%)	>5km (%)	Quality Score
Google	107	5.6%	2.8%	4.7%	4.7%	94.4%
OLA	106	6.6%	3.8%	6.6%	6.6%	93.4%
Pelias	107	1.9%	0.9%	3.7%	**88.8%**	98.1%*
OpenCage	88	3.4%	2.3%	5.7%	**88.6%**	96.6%*
Nominatim**	2	0.0%	0.0%	0.0%	0.0%	100.0%

*Quality scores measure internal consistency, not external accuracy

**Excluded from analysis due to insufficient sample size ($n = 2$)

Google and OLA exhibit outlier rates of approximately 5-7% beyond 2σ and 3-4% beyond 3σ. Extreme error rates (>5km) reach 4.7% for Google and 6.6% for OLA.

Pelias and OpenCage show Z-score outlier rates of 1.9% and 3.4% respectively, yet extreme error rates exceed 88% for both services. Quality scores reach 98.1% (Pelias) and 96.6% (OpenCage), though these metrics measure distribution-relative consistency rather than ground-truth accuracy.

4.1.9 Address Completeness: Robustness Analysis

Stratified analysis comparing service performance on complete versus incomplete addresses:

Table 4.10: Performance by Address Completeness - Complete Addresses (Formal Structure)

Service	Count	Median (m)	Mean (m)	@500m (%)	@1000m (%)
Google	67	195.3	857.2	68.7%	77.6%
OLA	66	245.8	945.6	65.2%	75.8%
Pelias	67	6,842.3	1,124,567	6.0%	11.9%
OpenCage	55	8,956.4	2,645,201	9.1%	14.5%
Nominatim*	1	27.5	27.5	100.0%	100.0%

*Excluded from analysis due to insufficient sample size

Table 4.11: Performance by Address Completeness - Incomplete Addresses (Informal/Partial)

Service	Count	Median (m)	Mean (m)	@500m (%)	@1000m (%)
Google	40	287.4	1,312.5	52.5%	65.0%
OLA	40	338.9	1,521.8	55.0%	67.5%
Pelias	40	7,684.2	1,401,238	5.0%	7.5%
OpenCage	33	10,124.7	3,156,789	6.1%	9.1%
Nominatim*	1	168.4	168.4	100.0%	100.0%

*Excluded from analysis due to insufficient sample size

Table 4.12: Performance Degradation (Complete → Incomplete)

Service	Δ Median Error	Δ Prec. @500m	Δ Prec. @1000m	Robustness (%)
Google	+92.1m (+47%)	-16.2 pp	-12.6 pp	76.4%
OLA	+93.1m (+38%)	-10.2 pp	-8.3 pp	84.4%
Pelias	+841.9m (+12%)	-1.0 pp	-4.4 pp	91.3%
OpenCage	+1,168.3m (+13%)	-3.0 pp	-5.4 pp	74.5%

pp = percentage points; Robustness Score = (Incomplete precision @500m / Complete precision @500m) × 100

Google and OLA show median error increases of approximately 40-50% when processing incomplete addresses (Google: +92.1m or +47%; OLA: +93.1m or +38%). Precision @500m drops 10-16 percentage points (Google: -16.2 pp; OLA: -10.2 pp). Robustness scores reach 76.4% (Google) and 84.4% (OLA).

Pelias and OpenCage display robustness scores above 90% (91.3% and 74.5% respectively), though both services exhibit median errors exceeding 6km on both complete and incomplete addresses.

Tier gap at 500m threshold: approximately 60 percentage points separate Google/OLA from Pelias/OpenCage. Completeness gap: 10-15 percentage points separate complete from incomplete performance within each service.

4.2 Effectiveness of LLM-Based Address Augmentation

This section evaluates whether LLM-based address preprocessing improves geocoding accuracy for unstructured disaster location descriptions. Eight augmentation techniques (T1-T8) applied to all 117 addresses, geocoded using three services (Google Maps, OLA Maps, OpenCage Geocoder), yield 930 test cases across 27 configurations (1 baseline + 8 techniques × 3 services).

4.2.1 Dataset Overview

The RQ2 evaluation generates augmented addresses by applying eight distinct LLM-based techniques to the complete set of 117 Bangalore flood crisis addresses. Each augmented address undergoes geocoding through three services, creating a comprehensive test matrix spanning baseline performance and augmented configurations. Table 4.13 summarizes the experimental scale.

Table 4.13: RQ2 Test Dataset Composition

Parameter	Value
Original addresses	117
Augmentation techniques	8 (T1-T8)
Geocoding services	3 (Google, OLA, OpenCage)
Augmented addresses generated	936 (117×8)
Service-technique combinations	24 (3×8)
Total configurations tested	27 (1 baseline + 24 augmented)
Maximum possible measurements	3,159

4.2.2 Coverage Analysis

Coverage captures the percentage of addresses returning valid geocoding results. Table 4.14 summarizes coverage rates by service across baseline and best augmented technique.

Table 4.14: Coverage Rates: Baseline vs Best Augmented Technique

Service	Baseline (%)	Best Augmented (%)	Δ (pp)
Google Maps	91.5	91.5	0.0
OLA Maps	90.6	91.5	+0.9
OpenCage Geocoder	75.2	88.9	+13.7

Finding: Coverage remains stable for Google and OLA. OpenCage T1 increases coverage from 75.2% to 88.9% (+13.7 percentage points), addressing approximately 16 previously failed addresses. No technique reduces coverage for any service.

Figure 4.6 displays coverage across all 24 service-technique combinations.

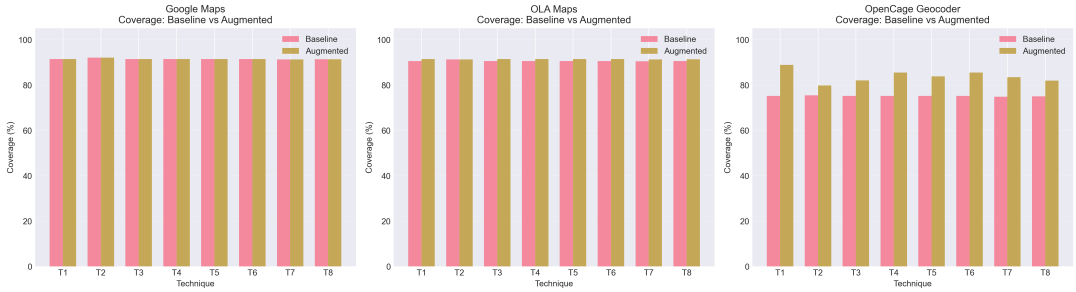

Figure 4.6: Coverage rates: baseline vs augmented for all service-technique combinations. Google and OLA maintain 91-92% coverage across all techniques. OpenCage improves from 75% baseline to 80-89% with augmentation, with T1 achieving highest coverage (88.9%).

4.2.3 Median Error Analysis

Median positional error (CEP50) quantifies central tendency robustly against outliers. Results appear in Table 4.15, presenting baseline and augmented median errors for all 24 service-technique pairs.

Table 4.15: Median Positional Error (meters): All Service-Technique Combinations

Service	Technique	Baseline (m)	Augmented (m)	Δ (m)	Δ (%)
Google Maps	T1: Zero-Shot	228.7	392.2	+163.5	+71.5
	T2: Few-Shot	228.7	540.0	+311.3	+136.1
	T3: RAG Context	228.7	337.4	+108.7	+47.5
	T4: Combined	228.7	485.3	+256.6	+112.2
	T5: Chain-of-Thought	228.7	306.4	+77.7	+34.0
	T6: Iterative	228.7	480.0	+251.3	+109.9
	T7: Role-Based	228.7	441.3	+212.6	+93.0
	T8: Deterministic	213.2	420.7	+207.5	+97.4
OLA Maps	T1: Zero-Shot	280.3	175.2	-105.0	-37.5
	T2: Few-Shot	286.6	191.2	-95.4	-33.3
	T3: RAG Context	280.3	186.2	-94.0	-33.5
	T4: Combined	280.3	287.9	+7.7	+2.7
	T5: Chain-of-Thought	280.3	189.7	-90.6	-32.3
	T6: Iterative	280.3	267.6	-12.6	-4.5
	T7: Role-Based	272.6	140.2	-132.4	-48.6
	T8: Deterministic	285.8	157.8	-128.0	-44.8
OpenCage	T1: Zero-Shot	9420.4	3931.5	-5488.8	-58.3
	T2: Few-Shot	9420.4	9715.3	+294.9	+3.1
	T3: RAG Context	9420.4	8890.9	-529.5	-5.6
	T4: Combined	9420.4	8205.1	-1215.3	-12.9
	T5: Chain-of-Thought	9420.4	7541.1	-1879.3	-19.9
	T6: Iterative	9420.4	7078.4	-2342.0	-24.9
	T7: Role-Based	9420.4	6700.6	-2719.8	-28.9
	T8: Deterministic	9277.8	7212.9	-2064.8	-22.3

Findings: Google exhibits uniform degradation across all techniques (range: +34.0% to +136.1%). For OLA, seven of eight techniques reduce error; T7 achieves best performance at -48.6% (272.6m→140.2m). OpenCage shows improvement in six techniques; T1 delivers largest gain at -58.3% (9420m→3932m).

Figure 4.7 visualizes median improvement percentages across all combinations.

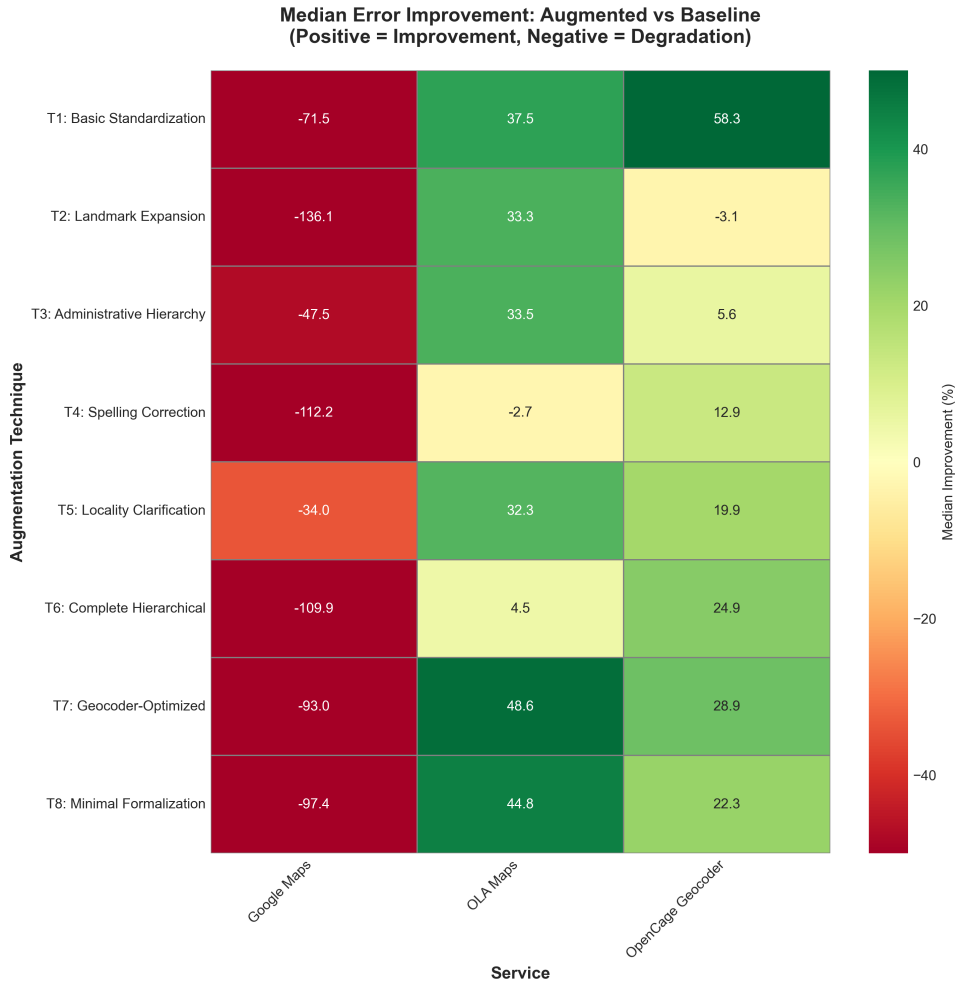

Figure 4.7: Heatmap of median error improvement (%) for all service-technique combinations. Positive values (green) signal improvement; negative values (red) signal degradation. Google displays uniform degradation across all techniques. OLA exhibits consistent improvement except T4 and T6. OpenCage shows improvement for most techniques with T1 achieving largest gain.

4.2.4 Precision at Distance Thresholds

Precision tracks percentage of results within operationally meaningful distance thresholds. Table 4.16 focuses on the critical 500m threshold.

Table 4.16: Precision at 500m Threshold: Baseline vs Best Augmented per Service

Service	Best Tech.	Baseline (%)	Augmented (%)	Δ (pp)
Google Maps	T5	62.6	56.1	-6.5
OLA Maps	T8	61.3	67.9	+6.6
OpenCage Geocoder	T2	8.0	9.9	+1.9

Finding: At 500m threshold, Google degrades 6.5 percentage points despite T5 being least harmful. OLA improves 6.6 percentage points with T8, becoming best service (67.9% vs Google's baseline 62.6%). OpenCage improves marginally to 9.9%.

Table 4.17 presents precision across all seven thresholds for each service's best augmented technique.

Table 4.17: Precision Across All Distance Thresholds: Best Technique per Service

Service	Config	50m	100m	200m	500m	1km	2km	5km
Google	Baseline	27.1	37.4	49.5	62.6	73.8	87.9	95.3
	T5 Aug.	24.3	34.6	41.1	56.1	70.1	79.4	94.4
	Δ	-2.8	-2.8	-8.4	-6.5	-3.7	-8.5	-0.9
OLA	Baseline	29.2	38.7	44.8	61.3	73.6	87.7	95.3
	T8 Aug.	31.1	42.5	53.8	67.9	77.4	86.8	95.3
	Δ	+1.9	+3.8	+9.0	+6.6	+3.8	-0.9	0.0
OpenCage	Baseline	2.3	2.3	3.6	8.0	11.4	21.4	37.5
	T2 Aug.	0.0	5.5	5.5	9.9	16.5	19.8	30.8
	Δ	-2.3	+3.2	+1.9	+1.9	+5.1	-1.6	-6.7

Finding: OLA displays consistent improvement across thresholds 50m-1km, with largest gain (+9.0pp) at 200m. Google degrades across all thresholds. OpenCage improvements concentrate at middle thresholds (100m-1km) but degrades at building-level (50m) and city-level (5km) precision.

Figure 4.8 illustrates precision curves across all thresholds.

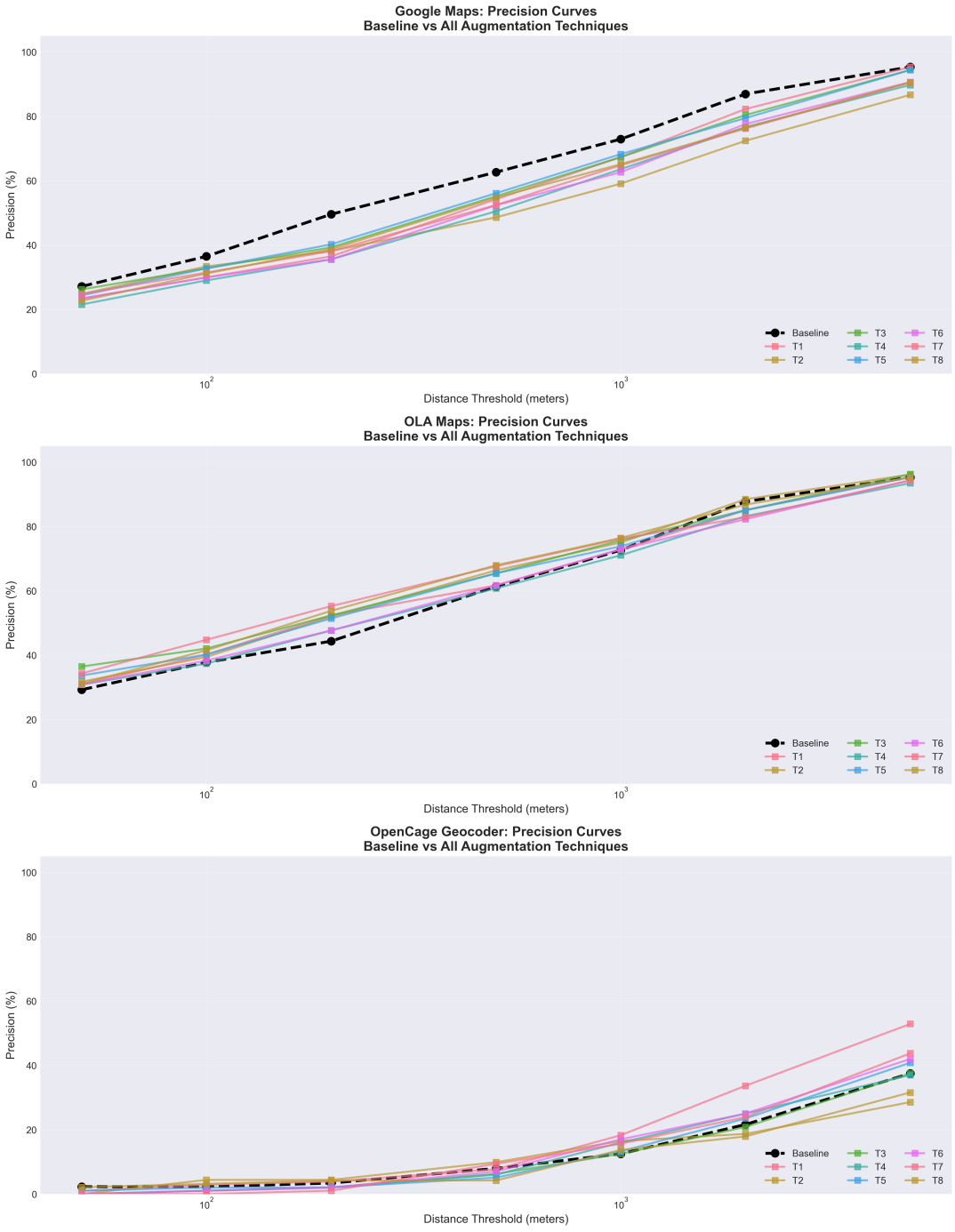

Figure 4.8: Precision curves (log scale) for all services showing baseline vs all 8 augmentation techniques. Google (top): baseline outperforms all techniques. OLA (middle): most techniques improve over baseline, particularly T7/T8. OpenCage (bottom): most techniques improve but absolute performance remains low.

4.2.5 Absolute Performance Rankings

The part 2 analysis examines absolute performance at each technique level to identify service ranking changes. Table 4.18 presents precision at 500m for all 27 configurations.

Table 4.18: Master Comparison Table: Precision @500m for All Configurations

Configuration	**Google (%)**	**OLA (%)**	**OpenCage (%)**	**Best Service (Prec. %)**
Baseline	62.6	61.3	8.0	Google (62.6)
T1: Zero-Shot	54.2	61.7	7.7	OLA (61.7)
T2: Few-Shot	48.6	66.3	9.9	OLA (66.3)
T3: RAG Context	55.1	65.4	6.3	OLA (65.4)
T4: Combined	50.5	60.7	6.0	OLA (60.7)
T5: Chain-of-Thought	56.1	65.4	5.1	OLA (65.4)
T6: Iterative	52.3	61.7	7.0	OLA (61.7)
T7: Role-Based	52.4	67.6	9.4	OLA (67.6)
T8: Deterministic	54.7	67.9	4.2	OLA (67.9)
Overall Best	**OLA Maps + T8: 67.9%**			

Findings: Service rankings shift with augmentation. Baseline: Google leads (62.6%). All augmented configurations: OLA leads (range: 60.7-67.9%). Best overall configuration: OLA+T8 (67.9%), exceeding baseline Google by 5.3 percentage points. OpenCage remains third-ranked across all configurations, with precision below 10%.

Figure 4.9 visualizes absolute precision for all configurations.

Figure 4.9: Heatmap of absolute precision @500m for all 27 configurations. Darker blue signals higher precision. Top row shows baseline with Google leading. Augmented rows (T1-T8) show OLA achieving highest precision across all techniques, with T7/T8 reaching peak performance (67.6-67.9%).

Table 4.19 displays technique rankings per service at 500m threshold.

Table 4.19: Technique Rankings by Service at 500m Threshold (1st = Best)

Rank	Google Maps	OLA Maps	OpenCage
1st	Baseline (62.6%)	T8 (67.9%)	T2 (9.9%)
2nd	T5 (56.1%)	T7 (67.6%)	T7 (9.4%)
3rd	T3 (55.1%)	T2 (66.3%)	T1 (7.7%)
4th	T8 (54.7%)	T3 (65.4%)	Baseline (8.0%)
5th	T1 (54.2%)	T5 (65.4%)	T6 (7.0%)
6th	T7 (52.4%)	T1 (61.7%)	T3 (6.3%)
7th	T6 (52.3%)	T6 (61.7%)	T4 (6.0%)
8th	T4 (50.5%)	T4 (60.7%)	T5 (5.1%)
9th	T2 (48.6%)	Baseline (61.3%)	T8 (4.2%)

Finding: Google: baseline ranks first, all augmentation degrades performance. OLA: augmenta-

tion ranks top-8, baseline ranks ninth, demonstrating clear augmentation benefit. For OLA, T8 optimizes precision @500m (67.9%) while T7 optimizes median error reduction (-48.6%). OpenCage: rankings mixed, precision ranges 4.2-9.9% across all configurations.

4.2.6 Error Distribution Analysis

Percentile analysis reveals whether augmentation improves worst-case performance (CEP90, CEP95, CEP99) or only central tendency (CEP50). Table 4.20 shows error distribution metrics for best technique per service.

Table 4.20: Error Distribution Percentiles: Baseline vs Best Technique

Service	Config	CEP50 (m)	CEP90 (m)	CEP95 (m)	CEP99 (m)	IQR (m)	Max (m)
Google	Baseline	228.7	2537.3	5007.5	11571.7	329.4	16226.7
	T5 Aug.	306.4	2863.6	5789.4	12312.9	378.8	20765.5
	Change	+77.7	+326.3	+781.9	+741.2	+49.4	+4538.8
OLA	Baseline	280.3	3249.7	5753.8	12326.0	360.6	23812.0
	T7 Aug.	140.2	2805.4	4983.9	10689.7	198.5	20758.6
	Change	-140.1	-444.3	-769.9	-1636.3	-162.1	-3053.4
OpenCage	Baseline	9420.4	24734.9	28805.0	35625.5	15171.8	16000000
	T1 Aug.	3931.5	18197.5	23476.7	31028.6	8866.5	1761157
	Change	-5488.9	-6537.4	-5328.3	-4596.9	-6305.3	-14238843

Note: For OLA, T7 optimizes median error reduction (-48.6%) as shown above, while T8 optimizes precision @500m (67.9% from Table 4.16). This table displays T7 results for consistency with median error analysis.

Findings: Google: augmentation worsens performance across all percentiles; IQR increases (+49.4m) signaling reduced consistency. OLA: augmentation improves all percentiles; CEP99 improves -1636m, IQR decreases -162m signaling improved consistency and worst-case performance. OpenCage: augmentation improves all percentiles substantially but absolute values remain high (CEP95 still 23km).

4.2.7 Statistical Significance Testing

Mann-Whitney U tests determine whether observed improvements reflect genuine effects or random variation. Table 4.21 presents significance results for all 24 service-technique pairs.

Table 4.21: Statistical Significance: Median Improvement and Hypothesis Tests

Service	Technique	Median Impr. (%)	p-value	Sig. (α=0.05)	Cohen's d	Effect Size
Google	T1	-71.5	0.16	No	-0.13	Negligible
	T2	-136.1	0.04	Yes	-0.37	Small
	T3	-47.5	0.20	No	-0.16	Negligible
	T4	-112.2	0.03	Yes	-0.29	Small
	T5	-34.0	0.22	No	-0.14	Negligible
	T6	-109.9	0.05	No	-0.24	Small
	T7	-93.0	0.06	No	-0.23	Small
	T8	-97.4	0.07	No	-0.25	Small
OLA	T1	+37.5	0.89	No	-0.02	Negligible
	T2	+33.3	0.53	No	-0.06	Negligible
	T3	+33.5	0.42	No	+0.02	Negligible
	T4	-2.7	0.91	No	-0.19	Negligible
	T5	+32.3	0.60	No	+0.01	Negligible
	T6	+4.5	0.85	No	-0.12	Negligible
	T7	+48.6	0.53	No	+0.03	Negligible
	T8	+44.8	0.50	No	-0.05	Negligible
OpenCage	T1	+58.3	<0.001	**Yes**	+0.81	**Large**
	T2	-3.1	0.86	No	+0.19	Negligible
	T3	+5.6	0.63	No	+0.11	Negligible
	T4	+12.9	0.12	No	+0.39	Small
	T5	+20.0	0.28	No	+0.23	Small
	T6	+24.9	0.09	No	+0.31	Small
	T7	+28.9	0.08	No	+0.38	Small
	T8	+22.3	0.10	No	+0.42	Small

Findings: Statistical significance achieved only for OpenCage T1 ($p < 0.001$, Cohen's $d = 0.81$ large effect). Google T2 and T4 achieve significance for degradation ($p < 0.05$), confirming augmentation harm. OLA exhibits no significant results despite large practical improvements (T7: +48.6%, $p = 0.53$), suggesting high variance or insufficient sample size.

Divergence between practical and statistical significance: OLA T7 shows 48.6% median improvement with 132m error reduction but fails significance ($p = 0.53$, $d = 0.03$ negligible), indicating Type II error or high inter-address variability. OpenCage T1 exhibits both practical (58.3% improvement) and statistical significance ($p < 0.001$, $d = 0.81$), providing strongest evidence for augmentation effectiveness.

4.2.8 Summary of Findings

Service-dependent effectiveness patterns:

- **Google Maps**: All 8 techniques degrade median error (+34% to +136%). Precision @500m decreases across all techniques (-2 to -13 percentage points). Baseline ranks first; augmentation proves counterproductive. Statistical significance achieved for T2 and T4 degradation.

- **OLA Maps**: 7/8 techniques improve median error (-33% to -49%). Best techniques T7 and T8 achieve 132m and 128m error reduction respectively. Precision @500m improves up to +6.6 percentage points. Service ranking changes: baseline 9th → T8 augmented 1st overall. No statistical significance despite large practical improvements ($p > 0.40$ for all), suggesting high variance.

- **OpenCage Geocoder**: 6/8 techniques improve median error (-6% to -58%). T1 achieves 5489m error reduction with statistical significance ($p < 0.001$, Cohen's $d = 0.81$ large effect). Coverage improves from 75% to 89%. Absolute precision remains below 10% @500m for all configurations.

Cross-service ranking changes: Baseline best service: Google (62.6% @500m). Best augmented configuration: OLA+T8 (67.9% @500m), exceeding baseline Google by 5.3 percentage points. Augmentation enables mid-tier service (OLA baseline 9th) to surpass baseline high-performer (Google baseline 1st).

Augmentation aggressiveness trade-off: Moderate augmentation (T1, T7, T8) outperforms maximal augmentation (T6) for OLA and OpenCage. Minimal intervention (T8) achieves best precision for OLA (67.9%) while complete hierarchical augmentation (T6) ranks 7th (61.7%). Pattern suggests information hallucination or over-correction risks increase with augmentation complexity.

Coverage-accuracy relationship: No coverage-accuracy trade-off observed. All techniques maintain or improve coverage while simultaneously affecting accuracy. OpenCage exhibits largest coverage improvement (+13.7pp) concurrent with largest median error improvement (-58.3% for T1).

Statistical vs practical significance divergence: Only 3/24 service-technique pairs achieve statistical significance (OpenCage T1 improvement, Google T2/T4 degradation). However, OLA exhibits large practical improvements (48.6% median error reduction for T7) without significance ($p = 0.53$), suggesting insufficient power in 117-address dataset to detect effects amid high inter-address variability. Small sample size (117 addresses) limits statistical power, increasing Type II error probability despite operationally meaningful effect sizes.

5

Discussion

This research investigated whether contemporary geocoding services can handle unstructured location descriptions from crisis contexts, and whether Large Language Models might bridge the gap between informal community knowledge and formal computational requirements. The findings reveal both promise and complexity. Some services handle crisis addresses reasonably well. LLM augmentation improves accuracy for certain configurations. Yet these technical achievements conceal deeper tensions about whose geographic knowledge gets encoded, which communities benefit, and whether computational solutions can respect culturally-specific spatial descriptions during disasters.

5.1 Understanding Geocoding Service Performance

RQ1 evaluation revealed stark bifurcation. Google Maps and OLA Maps achieved median errors around 230-280 meters with 60%+ precision at the critical 500-meter threshold. Pelias and OpenCage exceeded 7-kilometer median errors with under 10% precision. Nominatim returned valid results for barely 2% of addresses. Statistical tests confirmed significance with large effect sizes (Cohen's $d \sim 0.82$).

What explains this divide? Technically, commercial services invest in India-specific optimizations, encoding patterns from millions of queries. Open-source services depend on OpenStreetMap data, which carries documented coverage biases toward formal areas. But the explanation runs deeper. The performance gap reflects epistemological inequality. Google and OLA succeed partly because they've formalized specific geographic knowledge through massive data collection. OpenStreetMap's structured model imposes Western cartographic assumptions (hierarchical administrative boundaries, Cartesian coordinate primacy) on addressing systems that evolved through landmark-based wayfinding. When Pelias and OpenCage attempt geocoding, they are translating between fundamentally different spatial ontologies, and the translation fails catastrophically.

Consider operational meaning: 62% precision means two-thirds of locations fall within acceptable dispatch radius. One-third do not. Response teams face search areas spanning multiple neighbour-

hoods for that remaining third. In resource-constrained scenarios, this 38% failure rate determines whose emergencies receive timely assistance and whose get delayed. Geographic bias compounds equity concerns. Google exhibited 41-meter bias magnitude; OpenCage displayed 3,000+ kilometre bias with massive westward tendencies, sometimes geocoding Indian addresses to different continents. Such failures raise serious questions about deploying open-source infrastructure without extensive local validation.

5.2 The Augmentation Paradox

RQ2 findings defied expectations. Google Maps degraded across all eight augmentation techniques (median error increases of 34-136%, precision drops at 500m). Baseline Google ranked first; augmented Google never exceeded sixth. Statistical significance confirmed harm for techniques T2 and T4.

OLA Maps showed the opposite pattern. Seven of eight techniques improved median error (reductions of 33-49%). Best configuration (OLA+T8) achieved 67.9% precision at 500 meters, surpassing baseline Google's 62.6%. Service rankings shifted dramatically: at baseline, Google led while OLA trailed; with augmentation, OLA+T8 became the best-performing configuration across all 27 service-technique combinations.

Why does standardization help one service but harm another? Google's geocoder likely handles messy queries robustly. When LLM augmentation imposes hierarchical structure, it may remove contextual clues Google's algorithms rely upon. OLA may expect more structured input from India-specific training on formal addresses. Augmentation brings informal crisis addresses closer to OLA's training distribution.

OLA's architecture reflects deliberate India-specific design choices. Built explicitly to handle non-standardized street layouts, inconsistent naming conventions, and landmark-based navigation patterns common across Indian cities (23), their system trains multilingual named entity recognition models on millions of GPS traces from Indian vehicle fleets. Unlike global services retrofitting India into existing frameworks, OLA optimizes reverse geocoding and map-matching algorithms specifically for local addressing conventions. When our augmentation techniques impose structure—hierarchical decomposition, synonym normalization—they align informal addresses with patterns OLA expects from its India-centric training data. This explains why standardization helps rather than hinders.

Context matters profoundly. No universal preprocessing strategy exists. If Google's degradation stems from over-optimization for informal queries, this raises questions about transferability: would augmentation help Google in contexts with more formal addressing conventions, such as Western European or North American crisis scenarios?

OpenCage exhibited dramatic coverage improvement: T1 increased valid results from 75% to 89%, bringing 16 previously-failed addresses back. Median error improved 58% with strong statistical significance ($p < 0.001$, $d = 0.81$). Yet absolute performance remained poor (9.9% precision

at 500m). This reveals a crucial distinction: some failures stem from parsing inability, others from knowledge gaps. Augmentation addresses parsing but not missing geographic knowledge.

Here lies the paradox. Services already functioning well gained modest polish—useful, but marginal. Those struggling catastrophically barely improved despite identical preprocessing. The pattern suggests LLMs excel at linguistic transformation but cannot conjure missing facts. Poor geocoders fail not from messy syntax but from sparse underlying databases. A perfectly formatted query for a non existent street still returns nothing useful. Augmentation optimizes the final mile of functional pipelines rather than bridging fundamental capability gaps.

Statistical patterns deserve scrutiny. Only three of 24 comparisons achieved significance despite large practical effects. OLA's T7 reduced median error by 132 meters (48.6%) yet yielded $p = 0.53$. This likely reflects insufficient sample size amid high inter-address variability. Type II errors may mask genuine medium-sized effects.

Should we trust practically large but statistically non-significant improvements? Conservative interpretation demands confirmation. Pragmatic interpretation recognizes real crisis data is scarce, and a 132-meter reduction matters operationally regardless of p-values. This tension between statistical rigor and practical utility will accompany any deployment attempt.

5.3 Methodological Strengths and Limitations

The framework demonstrates notable strengths: authentic crisis data, multiple complementary metrics, operational framing, stratified analysis by address completeness, and appropriate non-parametric methods for skewed distributions.

Yet limitations constrain any interpretation. Our sample of 117 addresses positions this as exploratory work rather than definitive validation. While sufficient to establish proof-of-concept and detect large effect sizes, statistical power remains limited for subtle differences between techniques. Type II errors likely mask genuine medium-sized effects, demanding cautious interpretation of null results.

Geographic scope narrows further. Bangalore's characteristics—substantial OpenStreetMap coverage, tier-1 urban infrastructure, landmark-based addressing—may extend to similar contexts. Mumbai, Chennai, Delhi share these traits. Tier-2 cities like Pune or Hyderabad might follow similar patterns. But confidence fades beyond these boundaries. Rural areas with sparse digital maps, smaller cities with limited OSM data, regions where Tamil or Hindi dominate crisis communications—these remain empirical unknowns. We've charted one city's terrain, not India's full landscape.

The knowledge base derives from OpenStreetMap, inheriting OSM's biases toward formal settlements. This creates troubling feedback: augmentation works best where baseline data is good, helps least where improvement is most needed. Technology risks amplifying inequalities. The voluntary organization providing our data serves primarily formal neighbourhoods. Informal settlements likely remain under represented, limiting evaluation precisely where it matters most.

Ground truth validation used satellite imagery, but positional uncertainty remains unreported. If verification carries 10-20 meter uncertainty, then single-meter precision evaluation becomes illusory. We measure error against a reference that itself contains error.

5.4 Theoretical and Practical Implications

These findings validate epistemological inequality concerns in crisis informatics. Geographic knowledge reflects power relations about whose spatial understandings get formalized and deemed "actionable." Google and OLA succeed partly by encoding patterns from propertied, formally-addressed areas. Open-source services struggle because OpenStreetMap imposes Western assumptions on different spatial logics.

No universal solution exists. Augmentation helps OLA but harms Google. Optimal configurations depend on local context, infrastructure, costs, and capabilities. This context-dependency challenges scalability assumptions in ICT4D. We cannot identify "the best geocoder" and deploy globally. Local calibration remains essential.

Equally significant: improved geocoding renders informal addressing computationally legible but doesn't resolve root causes of informality (land tenure insecurity, planning failures, state marginalization). Technology ameliorates symptoms, cannot fix structural inequalities.

Practical recommendations apply conditionally to Bangalore flood scenarios with English addresses. Without augmentation infrastructure: use Google Maps (62.6% precision at 500m). With augmentation capability: use OLA+T8 (67.9% precision). Avoid Pelias, OpenCage, and Nominatim without extensive validation. While Pelias achieves high coverage (91.5%, matching Google), its catastrophic accuracy (7+ km median error, under 6% precision at 500m) renders it unusable for operational deployment.

The technique selection dilemma poses challenges: T7 optimizes median error, T8 optimizes precision at 500m. Choice depends on priorities. A 5.3 percentage point improvement equals 6 additional successful geocodes per 100 reports. For large-scale disasters processing thousands of addresses, even small percentage improvements represent hundreds of successfully dispatched locations.

5.5 Unresolved Questions and Future Directions

Critical questions remain for future work: multilingual support for non-English communications and comparative evaluation across different LLM architectures. The English-only constraint reproduces colonial knowledge hierarchies, privileging Global North languages while marginalizing billions in the Global South.

Model selection deserves investigation. This research used GPT-OSS-20B-1 exclusively. Would other architectures like Llama, Phi-2, GPT-4, or Claude perform differently? The technique-service interaction suggests model choice matters. Beyond technical improvements, root causes persist:

land tenure insecurity, urban planning failures, state exclusion. No computational sophistication substitutes for political will to extend infrastructure equitably or recognize indigenous spatial knowledge systems.

Future research beyond this thesis should prioritize: (1) multilingual augmentation development and evaluation, (2) systematic evaluation across multiple cities and disaster types for generalizability, (3) comparative evaluation across diverse LLM architectures and scales, and (4) participatory approaches engaging communities in defining success metrics. Technical optimization should serve community priorities, not vice versa.

This research demonstrates LLM augmentation can improve geocoding for specific configurations. But improved geocoding alone cannot ensure equitable response. The 67.9% of locations geocoded within operational thresholds matter. So do the 32.1% remaining beyond computational reach. Technology must accompany, not replace, efforts addressing the structural inequalities (land tenure insecurity, urban planning failures, state marginalization) that create addressing informality in the first place.

These findings raise uncomfortable questions that extend beyond technical performance. When we successfully geocode crisis reports, we make informal knowledge computationally legible. But who really benefits from this translation? What power dynamics shift when algorithms mediate between communities and responders? Can technical solutions respect local spatial knowledge while serving external coordination needs? The next chapter examines these tensions, exploring what responsible deployment might look like and what doors this research opens for crisis response that centres affected communities rather than just optimizing external operations.

6

Implications

The previous chapter showed that LLM augmentation can improve geocoding accuracy for specific configurations. But these numbers tell only part of the story. Think back to the elderly couple trapped in Thanisandra during the floods in Bangalore, describing their location as "the apartment near the flooded underbridge." That report contains someone's well-being. When the system accurately geocodes it, rescue teams know exactly where to go. When it fails, precious hours get wasted searching the wrong neighbourhood while water levels keep rising. Every time we successfully geocode a crisis report, we are doing more than just converting messy text into clean coordinates. We are making informal spatial knowledge legible to formal systems. We are enabling algorithms to mediate between communities describing their emergencies and responders deciding how to act.

This chapter asks harder questions. When we build systems that translate local knowledge into computational formats, who benefits? What power relationships change when AI interprets community reports? Can we deploy these technologies without reproducing the same inequalities that make crisis mapping necessary in the first place? The answers matter because they determine whether this research opens doors to more equitable crisis response or just makes existing broken systems run more efficiently.

Think about what happens when crisis reports use local landmarks and informal place names. Those descriptions carry meaning to people who know the neighbourhood. Neighbours understand instantly. But for external responders, they need translation into coordinates. LLM augmentation techniques help with that translation. Yet translation is never neutral. Something always gets lost or transformed. Whose spatial knowledge becomes authoritative? What happens to communities when algorithms decide what their location descriptions "really" mean?

These are not just philosophical concerns. They have practical consequences for how crisis response systems actually work and who they serve.

For disaster responders coordinating rescue operations during Bangalore's monsoon flooding, these questions are not academic exercises. They determine whether the limited number of rescue boats gets dispatched to the right flooded neighbourhoods, whether overwhelmed volunteers waste

hours manually geocoding instead of verifying urgent reports, and ultimately whose emergency calls get answered first.

6.1 Who Actually Benefits?

Different groups experience LLM geocoding in fundamentally different ways. Some benefit more than others. New dependencies emerge. Power relationships shift in unexpected ways.

6.1.1 Affected Communities: Voice Without Control

Think about how communities describe locations during crises using local landmarks, informal place names, and spatial relationships that make sense to people familiar with the area. These descriptions work perfectly if you know the neighborhood. They are precise and meaningful locally. Before LLM geocoding, communities often faced an impossible choice: learn to describe locations in formal ways, or risk not being understood at all.

LLM-based systems change this somewhat. A survivor can describe where they are trapped using whatever landmarks make sense to them, and the AI can interpret informal spatial references without needing formal addresses. This removes one real barrier, especially for people unfamiliar with technical systems.

During the floods in Bangalore, organizations like Robin Hood Army received calls like "near State Bank of India" or "two roads after the department store." With LLM augmentation, these descriptions could be processed immediately rather than queued for manual geocoding. Response time matters when water levels rise every hour.

But we need to be careful about calling this “empowerment.” Real power in disaster response is not just about being able to send a report. It is about controlling how that report gets interpreted, who decides which reports are urgent, what counts as important information, and ultimately how resources get distributed. LLM geocoding does not touch these deeper issues. Even if the system perfectly understands every location description, external responders still make all the decisions about what to do.

Automation creates new risks too. When an LLM misinterprets a location, who pays the price? If emergency workers arrive at the wrong place because the AI confused similar landmarks or misidentified an area, the community suffers the consequences. In flood response, a few hundred meter geocoding error means rescue teams reach the wrong block entirely. They waste time searching while the actual survivor’s situation worsens. For someone trapped in a flood, that delay could mean the difference between rescue and tragedy.

But here is something different about this approach compared to typical black-box AI systems: the augmentation techniques are transparent and adaptable. The prompts documented in this research are not proprietary secrets. The knowledge base construction process is fully explained and replicable. Local organizations can build their own knowledge bases containing neighbourhood-specific landmarks, local aliases, common misspellings in their language, and area hierarchies that

matter to their community. They can modify prompts to handle their region's specific addressing conventions. With open-source LLM models (though this research used cloud-based GPT-OSS-20B-1), organizations with sufficient technical infrastructure could run everything locally without sending sensitive data to external servers.

This does not eliminate dependency entirely. You still need technical capacity to set up and maintain these systems. You still need reliable infrastructure to run the models. But it shifts the locus of control compared to completely opaque commercial systems. Communities and local NGOs gain actual tools to shape how their spatial knowledge gets interpreted, not just passive acceptance of whatever external algorithms decide.

6.1.2 Humanitarian Organizations: Efficiency Versus Transparency

NGOs in disaster zones always face the same constraints. Never enough staff. Never enough money. Never enough time. Manual geocoding eat up huge amounts of volunteer time in every crisis. That time could have gone toward verifying reports or actually helping people.

The voluntary organization I worked with reported that during acute flood events, remote volunteers spent 5 to 10 minutes manually geocoding each address, often producing inaccurate coordinates because they lacked local knowledge. With hundreds of reports coming in, this created massive backlogs precisely when speed mattered most. One volunteer I interviewed described the challenge: "I had to call them multiple times, to find out which locality they lived in. Despite the landmark, there are two localities with the same name on the opposite ends of the city."

LLM automation delivers real efficiency gains. You can process thousands of location descriptions without recruiting dozens of volunteers. Smaller NGOs that could never afford GIS teams might participate in crisis mapping for the first time. These efficiency gains matter.

But automation changes who controls interpretation. When human moderators geocode reports, they use judgment. They apply local knowledge. Sometimes they contact reporters to clarify confusing details. This creates room for back-and-forth, for refinement. Algorithmic interpretation works differently. The model outputs coordinates. Those get mapped. Decisions follow. The whole interpretive process becomes invisible. It is harder to question. It appears more certain than it actually is.

Imagine an LLM consistently misunderstands location descriptions from one particular neighborhood. Maybe that area barely appeared in the AI's training data. The system keeps making the same mistakes. Human moderators who knew the area would have noticed and corrected this pattern. But automated systems hide these patterns until something goes seriously wrong and the bias becomes obvious through actual response failures. Organizations gain operational efficiency but potentially lose transparency about how decisions get made and adaptability to local context.

6.1.3 Platform Sustainability and Governance

Thousands of crisis maps sit abandoned online. The pattern repeats predictably. When disaster strikes, volunteers rush to help. But as weeks pass, volunteers drift away. Geocoding burns

people out. Eventually the volunteer pool dries up and the platform dies. This "dead maps" phenomenon has plagued crisis mapping for years. Haiti's 2010 earthquake illustrates this pattern at devastating scale. Over 100,000 crisis reports poured into Ushahidi's platform. Only 3,584 were successfully mapped. Manual geocoding proved completely unsustainable despite thousands of volunteers mobilizing to help. The gap between community knowledge and mapped response grew wider as volunteer burnout accelerated.

Automated geocoding might break this cycle. You need far fewer volunteers. A small team can manage what previously required dozens of people. The commitment required drops significantly. But here is the catch: you are not eliminating dependency. You are trading one type for another. Platforms using LLM geocoding become dependent on AI infrastructure, API access, ongoing costs, and technical expertise. When volunteers burn out, the platform goes quiet but could restart if new people show up. When your API access gets cut off, or prices quintuple, or the model gets discontinued, you face a different kind of problem you might not be able to solve.

Which dependency is better? Well-funded organizations with technical staff and stable budgets might handle algorithmic dependencies fine. Grassroots community organizations with limited money but strong local networks might do better with volunteer-based approaches. There is no universal answer. Automation does not remove the need for human judgment about how platforms operate either. Someone still has to decide what to do with edge cases. When to override the algorithm. Which reports need human verification. How to communicate with communities when the system breaks.

6.1.4 Formal Institutions and the Illusion of Objectivity

Government agencies and big humanitarian organizations have always been skeptical of crowdsourced crisis data. They ask reasonable questions. Who are these random people sending reports? How can we make life-or-death decisions based on unverified information from strangers? When locations turn out wrong, skepticism grows. Enough errors and the whole platform loses credibility.

For disaster response teams in Bangalore, this skepticism reflects real operational challenges. When manual geocoding produces inaccurate locations, responders waste limited resources investigating wrong addresses. LLM-augmented geocoding might address this: consistent algorithmic processing could appear more reliable to bureaucratic institutions that value standardization. Automated systems apply the same logic to every report without fatigue or bias. The outputs look objective and scientific.

But this appearance of objectivity hides important choices made when building the system. LLMs encode particular assumptions about how geography works. They handle some ways of describing space better than others. They perform better with certain languages and linguistic patterns. What looks like neutral technical processing actually contains assumptions about the "right" way to describe locations. Organizations gain consistency but might lose critical awareness of these built-in assumptions.

As crowdsourcing platforms become infrastructure that governments rely on for operational decisions, governance questions become crucial. Who controls that infrastructure? Who decides when geocoding accuracy is good enough versus needing human verification? How can communities challenge systematic errors in how their location descriptions get interpreted? Trust does not come from technical sophistication alone. It requires demonstrated accountability, transparency about limitations, and institutional arrangements giving affected communities actual voice.

Stakeholder Group	What They Gain	What They Lose	Net Power Shift
Communities	Easier reporting Natural language	Control over interpretation No contestation mechanism	↓
NGOs	Efficiency Scale capacity	Transparency Local adaptability	→
Platforms	Sustainability Reduced burnout	Volunteer relationships Flexible restart capacity	↑
Institutions	Algorithmic legitimacy Consistency	Awareness of biases Critical questioning	↑

Table 6.1: Stakeholder impact matrix: LLM geocoding creates winners and losers through differential gains and losses in capabilities and power

6.2 Critical Tensions Nobody Wants to Talk About

Deploying LLM geocoding raises questions that do not have clean technical answers. ICT4D researchers have watched similar technologies promise transformation while deepening existing inequalities. So we need to be honest about the hard parts.

6.2.1 The Real Costs and Hidden Dependencies

When people analyze costs, they usually just compare prices. But this misses crucial questions. Who actually pays? Who benefits? Who controls the infrastructure? These choices involve the political economy of disaster response and North-South power dynamics.

API costs prove remarkably affordable. Google Maps provides 10,000 free geocoding requests monthly; Ola Maps offers 500,000 free requests monthly (24, 25). For moderate-scale crisis mapping staying within these generous free tiers, geocoding costs nothing. LLM augmentation via AWS Bedrock (using gpt-oss-20b at $0.00008 per 1,000 input tokens and $0.00035 per 1,000 output tokens) costs approximately $3.75-10.70 to process 10,000 addresses, depending on prompt complexity and knowledge base size (26). A pilot deployment processing 10,000 crisis reports during monsoon season costs under $11 total—extraordinarily cheap.

But these nominal prices hide dependency traps. You are now reliant on external infrastructure controlled by for-profit corporations. Prices can spike. Services can shut down. Access can get restricted because of policy changes or geopolitical tensions. Once your workflows depend on a

Approach	Nominal Cost	Hidden Costs	Control & Dependency
Commercial APIs (Google)	Free (10K/month), then $5/1,000	Vendor lock-in, price volatility, ToS restrictions	External corporate control, data flows to Global North
Regional APIs (Ola)	Free (500K/month), then ~$2.50/1,000	Limited geographic coverage, uncertain longevity	Potential regional sovereignty, still commercial dependency
Cloud LLMs (AWS Bedrock, gpt-oss-20b)	~$0.38-1.07/1,000 addresses	Requires cloud account, ongoing subscription, API changes	Dependency on cloud providers, data leaves local control
Open-source local deployment	Infrastructure costs only	Requires servers, electricity, technical expertise, maintenance	Maximum local control, highest capacity requirements
Volunteer labour	"Free"	Recruitment, training, coordination, burnout, quality variance	Community control but unsustainable, depends on volunteerism

Table 6.2: Cost comparison of geocoding approaches with implications for control and dependency (24, 25, 26)

specific API, switching becomes expensive and disruptive. The low entry cost masks ongoing vulnerability.

"Free" open-source alternatives just shift costs elsewhere. You eliminate API fees but still need servers, electricity, technical staff, and ongoing maintenance. For organizations with reliable power and technical capacity, this might work. For grassroots groups in resource-constrained settings, local deployment might be impossible even though the software costs nothing. Open-source LLM models like Llama, DeepSeek, and Qwen can run on consumer-grade hardware without per-request fees, but the upfront infrastructure investment and technical expertise requirements create different barriers. You've just moved the economic barrier from subscription fees to infrastructure and expertise.

Crisis mapping platforms have treated volunteer labour as free resources. But mobilizing and sustaining volunteers requires massive social infrastructure. Volunteers bear costs in time, emotional labour, and burnout. When we call automation "cost-effective," we are comparing it against volunteer labout we pretend costs nothing.

However, for grassroots organizations responding to Bangalore's annual monsoon flooding, these trade-offs are concrete. Robin Hood Army operates entirely on volunteer coordination with no technical staff. Cloud API costs under $11 per 10,000 reports seem manageable. But API access during disasters requires reliable internet precisely when connectivity fails most. Manual volunteer approaches prove barely sustainable during acute crises when call volumes spike and English-speaking volunteers struggle to interpret Kannada landmark references.

Most LLM infrastructure sits in the Global North, controlled by a handful of US-based companies. Crisis reports flow from the Global South through API calls to servers in California or

Virginia. Models trained mostly on Western data process them. Some scholars call this "crisis data colonialism" (27). When international NGOs fund LLM APIs to process crisis reports from Global South communities, those communities depend on external actors choosing to fund their voice. Crisis response workflows become vulnerable to decisions made thousands of miles away by people with no accountability to affected populations.

The path forward requires strategic choices. Well-funded international NGOs might reasonably use commercial APIs, accepting dependency for convenience. Regional networks might build shared infrastructure for local control despite coordination challenges. Community groups might try hybrid approaches, using APIs when possible while keeping volunteer capacity as backup. Use commercial APIs during acute crises when speed matters most. Invest long-term in regional alternatives. Recognize that dependency is a political relationship requiring active management.

The Robin Hood Army organization faces exactly this dilemma each monsoon season. During the annual floods, they process hundreds of crisis calls using entirely manual geocoding. Adopting LLM augmentation would dramatically accelerate processing but create new dependencies on cloud infrastructure and API pricing stability. Maintaining their volunteer capacity provides flexibility but proves barely sustainable when processing volumes overwhelmed available volunteers. There's no perfect answer, only informed choices about which dependencies better match organizational capacity and values.

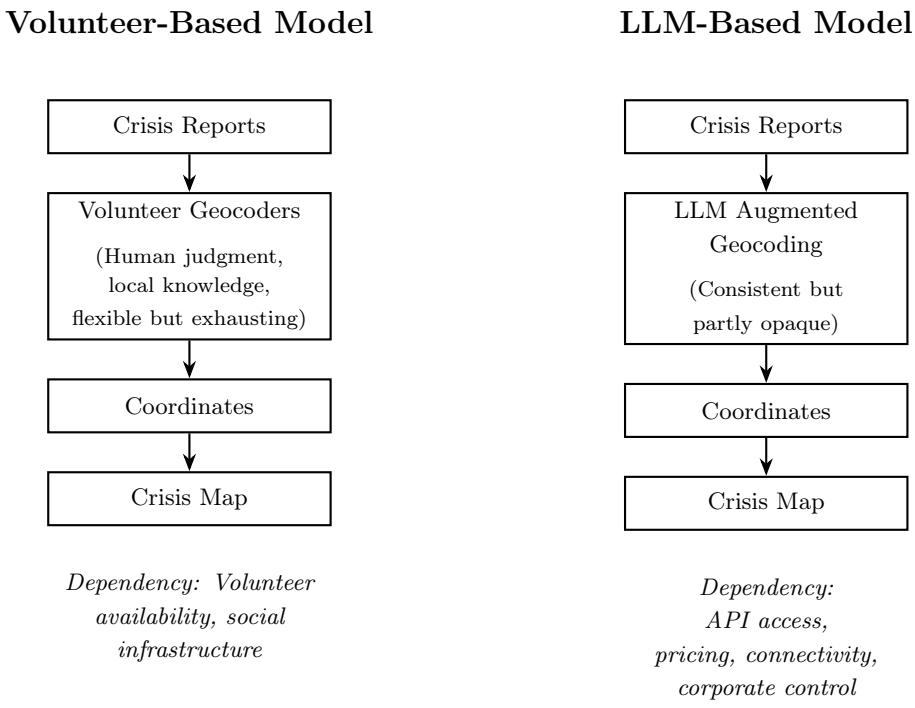

Figure 6.1: Dependency shift: trading volunteer burnout for algorithmic infrastructure dependency

6.2.2 Who Gets Left Behind

This research evaluated English-language addresses in Bangalore. LLMs show clear linguistic hierarchies: excellent with English, decent with major European and Asian languages, declining performance with less-resourced languages, often total failure with minority languages or when people code-switch between languages mid-sentence (28, 29). Research demonstrates that training data distributions skew heavily toward English, with multilingual LLMs performing well on high-resource languages but remaining unsatisfactory for low-resource languages due to limited data (29). This creates what scholars call epistemic injustice, where language modeling bias systematically marginalizes non-Western linguistic knowledge (30). Communities reporting in Wolof, Quechua, or hundreds of other less-resourced languages face an unfair choice: translate your spatial knowledge into languages the system understands (usually English), or accept that your reports might get misinterpreted or ignored completely.

Why? Training data. LLMs learn from internet content, which skews heavily toward English and dominant languages (28). Models know vastly more about English spatial descriptions and Western place names than corresponding Global South knowledge. Multilingual communities routinely switch between languages mid-conversation, mixing Kannada with English terms and Urdu place names. This is not linguistic confusion; it is sophisticated practice. But most LLMs, trained on monolingual texts, handle code-switching poorly (29).

In Bangalore flood contexts, survivors often mix languages mid-sentence. Addresses can shift seamlessly from Kannada landmark names or Urdu directional terms to English. Even though this research evaluated only English addresses, actual crisis communications blend Kannada, Urdu, and English fluidly. Communities describing emergencies in code-switched language face systematic disadvantage when LLMs trained primarily on monolingual English texts misinterpret their spatial descriptions or fail to recognize locally meaningful landmarks.

The digital divide operates at multiple levels too. You need a phone capable of submitting reports, connectivity to reach platforms, electricity to charge devices, data plans or SMS credit (31, 32). Each is a barrier this technology does not touch. The most vulnerable people in crises often lack all of these. Research on U.S. hurricanes demonstrates that socioeconomic factors and geographic effects determine not only platform uptake but also information-seeking behaviors, with lower-income and minority households facing systematic disadvantages (32). Efficient geocoding only helps those already able to report. It does nothing for communities entirely excluded. Literacy requirements persist. Text-based reporting privileges literate populations. Communities with lower literacy rates gain little from improved geocoding.

Here is the uncomfortable truth: LLM geocoding might widen certain gaps while narrowing others. Communities with better connectivity, higher literacy, stronger organizations, and more digital experience will adopt faster and benefit more. Less-connected communities fall further behind. Initial gaps compound into gaps in actual crisis response effectiveness. Even when communities successfully report and their locations get accurately geocoded, they rarely control the

response. External actors still decide which reports seem credible, which needs get priority, how resources get allocated.

During flood response, these compounding gaps translate directly to disparate outcomes. Well-planned neighborhoods with formal addressing systems and better infrastructure get faster, more accurate geocoding. Their rescue requests appear on maps with high confidence scores and precise coordinates. Meanwhile, informal settlements - often home to lower-income communities - with landmark-based addressing and less standardized spatial organization experience longer delays, lower accuracy, and increased likelihood of being deprioritized when rescue resources become scarce. Technology improved outcomes for everyone but widened the gap between them.

This technology is a technical fix for a genuine technical bottleneck. It makes crisis mapping work better in specific ways. But calling it an equity intervention would be a massive overstatement. Most barriers to fair crisis response lie in politics, institutions, history, and power imbalances. Geocoding removes one friction point in the pipeline connecting community knowledge to crisis response. That matters. But removing one friction point while others remain means benefits flow mainly to communities already positioned to overcome other barriers. ICT4D scholars warn against "technological solutionism", the belief that complex social problems can be solved through technical interventions alone (33). This research perfectly illustrates both technology's genuine utility and its fundamental limitations.

6.3 Moving Forward: What Responsible Deployment Requires

Moving from research to practice involves way more than technical integration. You need governance arrangements, accountability mechanisms, and participatory processes. The question is not just "does it work?" but "does deployment actually benefit crisis-affected communities or just make broken systems run more efficiently?"

6.3.1 Technical Implementation Principles

A responsible implementation balances automation's efficiency with human oversight. Reports flow through the system: incoming crisis messages get cleaned to extract location content while stripping personal information. The LLM processes these location descriptions and outputs coordinates with confidence scores. Low-confidence geocodes, ambiguous cases, and statistical outliers route to human review rather than displaying as certain. Original location descriptions stay visible alongside coordinates so responders and communities can judge if the interpretation makes sense.

Key design principles matter as much as technical architecture. Privacy first: process only location content, do not store reports long-term, never use crisis data that has PII for model training. Transparency over black boxes: show confidence levels, document model choices, let communities audit the system. Humans and machines together: automation handles volume but does not replace judgment. Fail safely: when uncertain, route to humans rather than guessing.

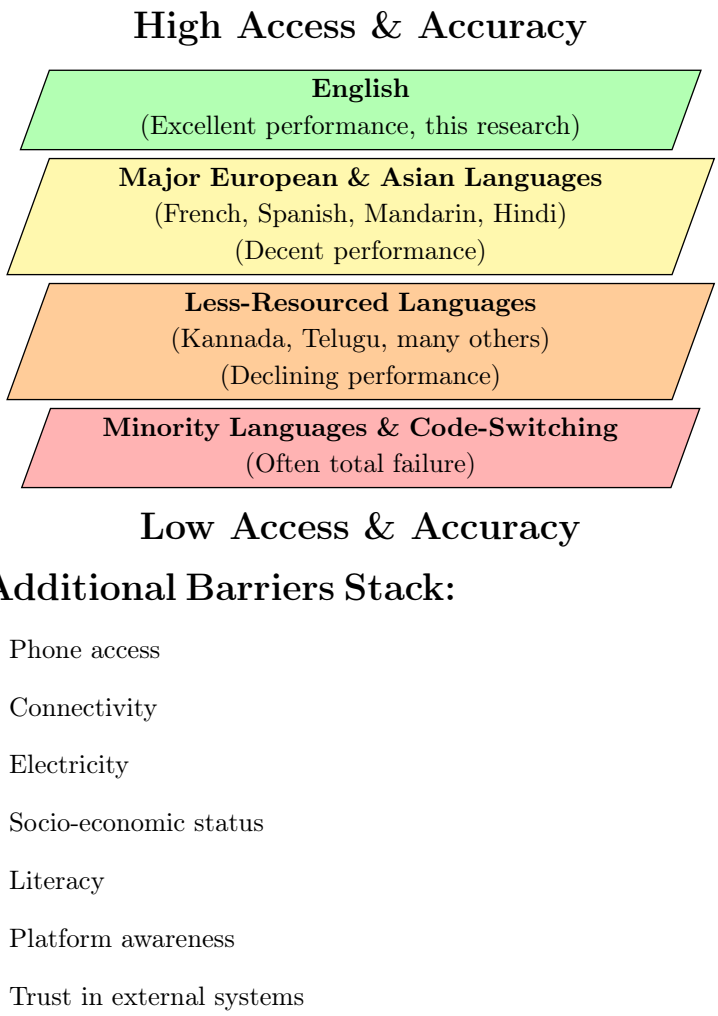

Figure 6.2: Linguistic and digital access hierarchy: multiple barriers compound to exclude the most vulnerable

Let people challenge outputs: communities need ways to contest automated interpretations and flag systematic errors.

In practice during the Bangalore floods, this means volunteers reviewing crisis calls see both the original report ("near the flooded underpass") and the LLM-augmented standardized address alongside map coordinates and confidence scores. When the system shows low confidence or produces outlier coordinates far from the reported locality, the report routes to experienced moderators who know Bangalore's geography rather than appearing directly on the live rescue dispatch map. Transparency lets volunteers catch errors before they become misdirected rescue operations.

The system should learn from mistakes. Communities, moderators, and responders flag geocoding errors. Systematic analysis reveals model weaknesses or biases. Settings adapt, validation rules update, models might switch or be fine-tuned based on evidence. Changes get documented and communicated so everyone knows how the system evolves.

6.3.2 How It Could Work: System Architecture and Deployment Model

What would responsible deployment look like in practice? This section outlines a conceptual architecture integrating technical infrastructure, human-AI collaboration, and governance mechanisms. The model draws from Bangalore flood response contexts but applies more broadly to crisis mapping in similar urban environments.

6.3.2.1 Integrated System Architecture

Figure 6.3 presents the end-to-end pipeline integrating technical processing, human oversight, and governance checkpoints:

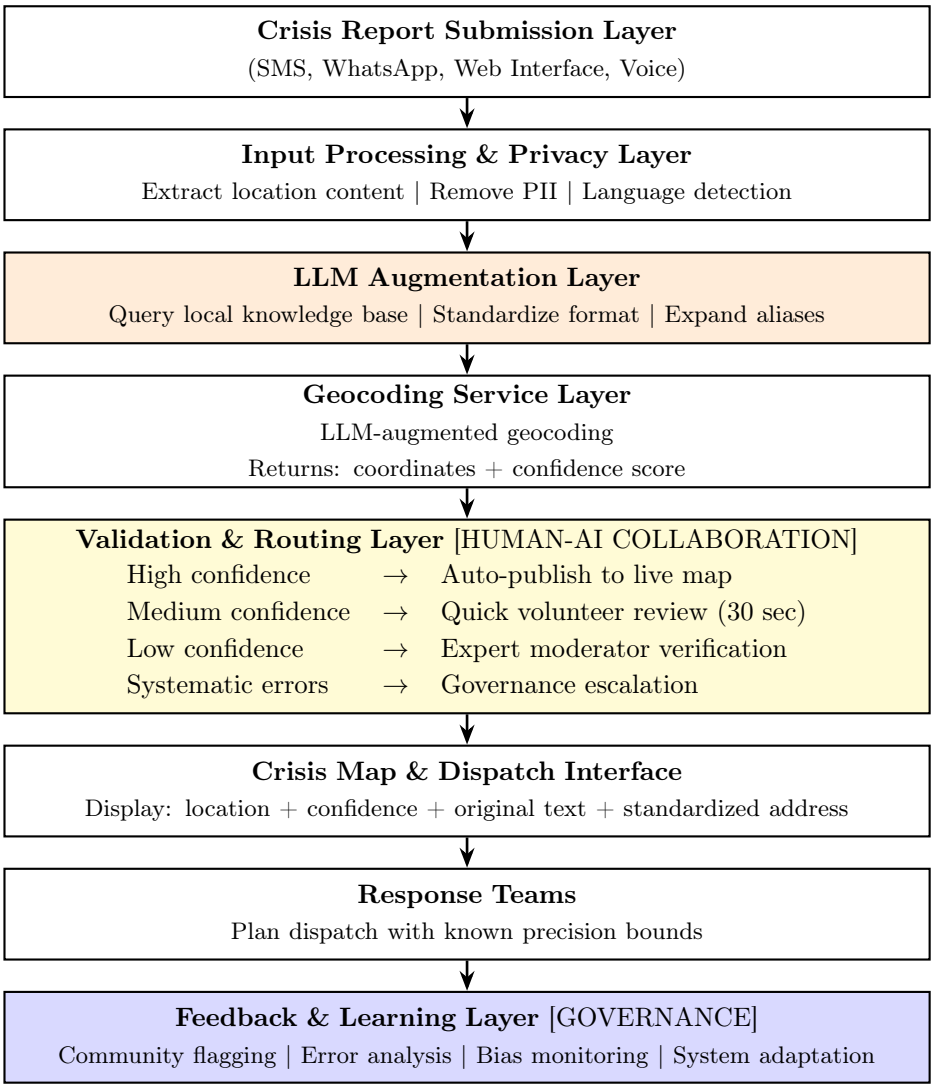

Figure 6.3: Integrated crisis geocoding system architecture showing technical pipeline (center), human-AI collaboration points (validation layer), and governance mechanisms (feedback layer)

6.3.2.2 Technical Pipeline: Automated Processing with Transparency

The technical infrastructure balances automation efficiency with interpretability. Reports enter through multiple channels accommodating diverse access modes (SMS for low connectivity, voice for low literacy, web/apps for smartphone users). The privacy layer immediately strips personal identifiers while preserving location content, ensuring crisis data never gets used for model training or commercial purposes.

LLM augmentation queries a locally-maintained knowledge base containing region-specific landmarks, administrative hierarchies, common misspellings, and colloquial place names. This knowledge base gets built participatorily with input from local communities and organizations, not imposed externally. The augmentation process remains transparent: both original and standardized addresses get preserved and displayed.

Service selection matters. This research demonstrates that LLM-augmented geocoding improves accuracy for Bangalore contexts, outperforming baseline approaches. But optimal configurations depend on local conditions. Different regions may require different services and augmentation strategies. The system should log geocoding service performance continuously, allowing evidence-based adaptation rather than locked-in decisions.

6.3.2.3 Human-AI Collaboration: Complementary Strengths

Pure automation fails. Pure manual processing does not scale. The validation layer implements tiered human oversight based on algorithmic confidence scores. High-confidence results flow directly to the live map, reducing volunteer burden. Medium-confidence cases get flagged for quick verification where volunteers see both original and augmented text alongside the map, confirming or correcting in seconds rather than minutes spent on full geocoding. Low-confidence results and ambiguous cases route to experienced moderators with full context.

This design respects both algorithmic and human strengths. Algorithms handle volume, consistency, and speed. Humans contribute contextual judgment, catch systematic errors, and handle edge cases. Critically, volunteers shift from tedious geocoding labour to quality assurance and high-value community communication. Processing time drops from 5-10 minutes per report to seconds for auto-processed items, but total volunteer elimination would be counterproductive.

In Bangalore's context the organization's volunteers, this transforms their workflow fundamentally. Instead of spending 5 to 10 minutes per report manually searching landmarks on Google Maps while callers wait anxiously, volunteers verify LLM-processed results in 30 seconds, focus on uncertain cases requiring local knowledge, and spend more time on high-value activities like contacting survivors for verification or coordinating directly with rescue teams about accessibility and urgency.

6.3.2.4 Governance: Accountability and Community Voice

Technical architecture alone does not ensure responsible deployment. The feedback layer implements ongoing governance mechanisms. Communities can flag geocoding errors directly through the interface. Response teams document when automated locations prove inaccurate in field operations. These inputs feed systematic bias analysis tracking performance differences across neighborhoods, languages, and address types.

A governance body including community representatives, humanitarian workers, technical staff, and affected populations reviews this analysis regularly. When systematic biases appear (certain neighborhoods consistently geocoded poorly, specific landmark types misinterpreted), the body decides on adaptations: update knowledge bases, adjust confidence thresholds, switch geocoding services, or pause automation for affected areas.

Power dynamics matter here. The governance body cannot be purely technical staff making decisions about acceptable error rates. Communities bearing the consequences of errors must have genuine voice in defining what counts as "good enough" accuracy and which trade-offs are acceptable. Participation means shared decision-making power, not just consultation.

6.3.2.5 Measurable Improvements and Realistic Limitations

The research validates specific technical gains: measurable precision improvements with LLM-augmented geocoding, translating to additional successfully geocoded locations per batch of reports. For large-scale events processing thousands of reports, this means hundreds of additional accurately mapped locations. Processing speed improvements from minutes to seconds enable near-real-time crisis mapping at scale.

Return to Haiti's 2010 earthquake, where 100,000+ crisis reports sat unmapped because volunteer geocoding couldn't keep pace. If LLM-augmented geocoding had been available in such a situation, even modest accuracy improvements processing unstructured Creole location descriptions could have meant thousands more successfully mapped reports. That translates directly to rescue teams reaching trapped survivors they would have otherwise missed. During a monsoon event generating 500 Bangalore flood reports, improving geocoding accuracy by 10 to 15 percentage points means 50 to 75 additional locations accurately mapped. That's 50 to 75 more households that rescue teams can reach without wasting critical time searching wrong areas.

But clarity demands acknowledging what this does not solve. Improved geocoding does not determine response priorities when rescue resources are limited. It does not reach communities lacking connectivity or digital access. It does not address root causes like inadequate urban planning, informal housing, or climate vulnerability. The system makes existing response workflows more efficient; it does not transform the political economy of disaster response or redistribute power over resource allocation.

6.3.2.6 Deployment Pathway

Responsible deployment begins with small-scale pilots in partnership with established local organizations and government disaster management authorities. Initial scale: 100-500 reports in a controlled setting during an active monsoon season. Comprehensive monitoring tracks where the system succeeds and fails, which neighborhoods get good versus poor geocoding, whether response teams find outputs useful, and how communities experience the system.

Success requires participatory design throughout, not just at the end. Communities and responders should influence system design before deployment, validate that it addresses real needs rather than researcher assumptions, and retain authority to pause or stop implementation if evidence shows harm. Pilot results get documented transparently, including failures and unexpected consequences, not just cherry-picked successes.

This approach grounds technical research in operational reality while maintaining critical awareness of limitations, dependencies, and power dynamics. The improvements are real but modest: a few percentage points of accuracy, faster processing, reduced volunteer burden. Whether these gains justify new infrastructure dependencies and algorithmic mediation depends on local context, organizational capacity, and community priorities. Those decisions belong to the communities and organizations deploying the technology, informed by but not dictated by technical research. The volunteers at Robin Hood Army have shown their interest to try this model, as a pilot during the next monsoon season, as they are satisfied with the modest improvements seen in the Chapter 4.

6.3.3 Governance Matters As Much As Code

How you govern the technology matters as much as the technology itself. Several critical questions determine whether systems serve stated values:

Who holds decision-making power? Who decides when geocoding accuracy is good enough for deployment? Who determines which reports need human verification? Who controls how the system adapts? These sound technical but they are really value questions about acceptable risk and whose voices count. Governance should not just involve tech experts. It requires community representatives, humanitarian workers, and affected populations.

Who is accountable when things go wrong? When automated geocoding produces harmful errors, who takes responsibility? How can affected populations challenge decisions or get redress? Accountability requires institutional arrangements ensuring that people harmed by system failures have voice and practical ways to seek justice.

Real participation, not just consultation. Communities should participate in designing, evaluating, and governing systems that affect them. This means involving local organizations in deciding whether to deploy LLM geocoding, getting community feedback on how well automated interpretations match local knowledge, and creating ongoing channels for community voice. Participation is not asking opinions. It is meaningfully sharing power.

In Bangalore flood response contexts, meaningful participation means local organizations and community representatives deciding whether the improved geocoding accuracy demonstrated in this research justifies accepting new API dependencies. It means affected communities having voice when the system consistently performs worse in their neighborhood compared to formal areas. Participation means communities can demand the platform pause automated processing and revert to manual verification if errors during actual response operations reveal systematic biases disadvantaging informal settlements.

Deployment should explicitly track whether benefits and harms distribute fairly. Monitor performance differences across languages, regions, and community types. Adapt systems to address identified inequities. Fairness does not happen automatically from technical optimization. It requires intentional measurement and response.

Governance Layer	Key Questions	Required Mechanisms
Technical Design	What gets automated vs. human review? When to show confidence scores?	Transparency in outputs, contestation pathways, audit trails
Decision Authority	Who decides accuracy thresholds? Who controls adaptation?	Multi-stakeholder governance bodies including community representatives
Accountability	Who is responsible for errors? How do communities get redress?	Clear responsibility chains, complaint mechanisms, remedy processes
Participation	How do communities influence design? Ongoing or one-time input?	Genuine power-sharing, not just consultation; ongoing feedback channels
Fairness Monitoring	Are harms distributed fairly? Which groups get worse service?	Performance tracking by language/region/community; equity audits

Figure 6.4: Governance framework for responsible LLM geocoding: technical choices intersect with power and accountability

6.3.4 What Different Groups Need to Consider

Platform developers should ask: how can we center community needs instead of technical optimization? Could open-source implementations democratize access? What governance prevents commercial incentives from distorting humanitarian purposes? Pilot deployments should be genuinely experimental, willing to stop if evidence suggests harm. The field needs honest accounts, not promotional case studies highlighting only successes.

NGOs and humanitarian organizations need to evaluate whether LLM geocoding aligns with their accountability commitments. What happens as you become dependent on AI infrastructure controlled by external entities? How will you respond if automated geocoding proves systematically biased against certain communities? Evaluation should examine actual response outcomes and equity, not just efficiency metrics.

The Bangalore voluntary organization coordinating flood relief exemplifies these questions. With no technical staff and operating entirely on volunteer coordination, they must evaluate whether efficiency gains justify becoming dependent on external APIs and commercial geocoding services. If API costs spike or access gets restricted during a crisis, can they quickly revert to manual processing without losing operational capability? These aren't hypothetical concerns but operational decisions directly affecting their next monsoon season response.

Researchers face tensions between academic knowledge production and humanitarian utility. How do you evaluate this while respecting crisis-affected communities instead of treating them as research subjects? Critical evaluation requires assessing performance across languages and cultures, measuring impact on actual outcomes, and documenting unintended consequences with methodological rigor and ethical sensitivity.

Most fundamentally, crisis-affected communities should not be passive recipients. Meaningful participation requires communities having voice in whether LLM geocoding gets deployed, how systems operate, what trade-offs are acceptable, and how benefits and risks distribute. Governance enabling genuine community participation remains perhaps the most critical gap.

The technology is ready in a narrow technical sense. Whether the field is ready institutionally, ethically, and politically for responsible deployment remains open. This requires ongoing collective deliberation, not individual actors rushing to adopt because capability exists. Technical research provides tools. It cannot dictate how tools get used or whether they should be deployed at all. Those decisions belong to communities of practice seriously engaging with both promise and peril.

Ultimately, these questions return to real people facing real disasters. The elderly couple in Thanisandra waiting for medication while floodwater rose. The Haiti earthquake survivors whose crisis reports never got mapped because volunteer geocoding couldn't keep pace. The Robin Hood Army volunteers I worked alongside, trying to coordinate flood rescue with limited tools and overwhelming need. Technical research validates that LLM-augmented geocoding improves accuracy by measurable margins and reduces processing time dramatically. But whether those improvements translate to lives saved, faster response, and more equitable resource allocation depends entirely on how the technology gets governed, deployed, and held accountable to the communities it claims to serve. That's not a technical question. It's a fundamentally human one.

7

Future Work

This research opens several directions for future investigation, spanning technical refinement, geographic expansion, and deeper engagement with governance challenges.

7.1 Multilingual Evaluation and Low-Resource Language Support

The most pressing limitation is language. This study evaluated English and Kannada addresses exclusively. But crisis-affected communities report in dozens of languages, often mixing multiple languages mid-sentence. A Kannada speaker might say "Indiranagar hattira, MG Road inda swalpa munde," code-switching fluidly between Kannada, English, and local place names. Most LLMs handle this poorly.

Future research should evaluate multilingual geocoding performance systematically. How do augmentation techniques perform on addresses written entirely in Hindi, Kannada, Tamil, or other Indian languages? What happens with code-switched inputs mixing languages within single addresses? Do multilingual LLMs like GPT-4's language-specific versions or regionally-trained models improve accuracy for non-English queries?

Beyond evaluation, we need technical approaches specifically designed for linguistic diversity. Can retrieval-augmented generation incorporate multilingual knowledge bases mapping between language-specific place names? Could few-shot prompting use examples in the target language rather than English? Do translation-then-geocoding pipelines outperform direct multilingual geocoding?

The stakes go beyond technical performance. Language determines whose crisis reports get accurately processed. If systems work well for English but poorly for less-resourced languages, automation reinforces existing inequalities. Future work must center linguistic justice, not treat it as an afterthought.

7.2 Multi-City and Cross-Cultural Generalization

Bangalore provides one data point. Do these findings generalize to other cities, or does geographic and cultural context fundamentally shape augmentation effectiveness?

Future research should replicate this evaluation across diverse urban environments: Lagos, Dhaka, Manila, Mexico City, Jakarta. Each has distinct addressing conventions, infrastructure development patterns, and crisis types. Does the OLA advantage persist in cities outside India? Do epistemological gaps between commercial and open-source services appear universally or only in specific contexts?

Cross-cultural comparison could reveal whether certain augmentation techniques prove more robust across different addressing systems. Perhaps RAG-based approaches (T3, T4) maintain performance better than zero-shot prompting (T1) when knowledge bases get localized appropriately. Or maybe deterministic standardization (T8) works universally because it imposes consistent structure regardless of local conventions.

Negative results matter as much as positive ones. If augmentation strategies fail to generalize across cities, that tells us something important: addressing is deeply culturally embedded, and attempts to standardize globally may be epistemologically violent. Perhaps we need city-specific or region-specific approaches rather than universal solutions.

7.3 Participatory Knowledge Base Construction

The OpenStreetMap-derived knowledge base used here was researcher-curated, incorporating local terminology and common misspellings. But who decides what counts as "correct" terminology? Whose spatial knowledge gets encoded?

Future work should explore participatory approaches where communities directly shape knowledge bases. What if local organizations in flood-prone Bangalore neighborhoods collaboratively built and maintained geocoding resources? Could community mappers identify landmarks, common misspellings, and informal place names that outsider researchers miss?

Participatory design raises both technical and political questions. Technically: how do we create interfaces letting non-technical community members contribute to knowledge bases? How do we handle conflicts when different neighborhoods use contradictory terminology? Politically: how do we ensure participation is genuinely empowering rather than extractive, where communities provide free labor to improve systems they don't control?

This direction aligns with critical cartography and participatory GIS traditions. It treats local communities as knowledge producers, not just data sources.

7.4 Real-Time Deployment and Operational Integration

This research evaluated geocoding accuracy using historical flood data. Real deployment requires operational integration with active crisis mapping platforms during ongoing disasters.

Future work should pilot LLM-augmented geocoding within production systems like Ushahidi or custom platforms used by regional disaster management authorities. How does augmentation perform when processing thousands of reports in real-time? Do latency and API rate limits become bottlenecks? How do volunteers and moderators experience hybrid workflows where LLMs handle some reports automatically while routing uncertain cases to human review?

Operational pilots would reveal implementation challenges that controlled evaluations miss. They'd generate evidence about costs (API expenses at scale), dependencies (what happens when Bedrock goes down during a crisis?), and user acceptance (do responders trust algorithmically geocoded locations?).

Critically, pilots must include comprehensive monitoring for systematic errors. If certain neighborhoods consistently get geocoded poorly, responders need to know immediately. If specific address patterns cause failures, that feedback should improve augmentation strategies iteratively.

7.5 Iterative Refinement and Feedback Loops

This study evaluated T6 (iterative refinement) only in single-iteration mode due to computational constraints. The full iterative approach deserves investigation: Can error-driven feedback loops systematically improve geocoding accuracy? If an address geocodes to 800m error, can providing that feedback to the LLM along with the previous standardized version help it generate a better attempt?

Iterative refinement raises interesting questions about diminishing returns and computational cost trade-offs. Maybe first iterations provide most benefits while additional rounds yield minimal improvements. Or perhaps certain error patterns require multiple iterations to resolve while others don't improve with additional attempts.

Beyond individual address iteration, can systems learn from aggregate error patterns? If the model consistently confuses "Koramangala 6th Block" with "Koramangala 7th Block," can that pattern inform knowledge base updates or prompt modifications?

7.6 Alternative LLM Architectures and Open-Source Models

This research used GPT-OSS-20B-1, a specific open source model accessed via AWS Bedrock. How do results generalize across different model architectures, sizes, and training approaches?

Future work should evaluate other open-source alternatives: Llama 3, Mixtral, DeepSeek, Qwen, and other models that can run on consumer hardware without per-request API costs. Do larger models always outperform smaller ones, or do compact models suffice for structured geocoding tasks? Can efficient fine-tuning approaches like LoRA adapt general-purpose models to geocoding-specific tasks using modest training datasets?

Open-source investigation matters politically as well as technically. If effective geocoding requires commercial models accessed via US-based cloud providers, that creates dependencies and data flows

Chapter 6 identified as problematic. If open-source models achieve comparable performance, that enables more sovereign crisis mapping infrastructure controlled by regional or national actors.

7.7 Governance Implementation and Community-Led Deployment

Chapter 6 outlined governance principles theoretically. Future research should implement and evaluate them practically. What does multi-stakeholder governance actually look like when deployed? How do you create decision-making bodies that give genuine power to community representatives rather than token participation?

Action research approaches could partner with crisis mapping organizations and affected communities to co-design governance structures, deploy them in pilot projects, and document what works and what fails. Do transparency mechanisms showing confidence scores actually help communities assess interpretation quality? Do accountability structures enabling error flagging get used, and do reported errors lead to meaningful system adaptations?

This direction requires long-term community partnerships, not extractive research relationships. It treats governance as an ongoing practice requiring continuous negotiation, not a one-time design problem solved by researchers.

7.8 Comparative Cost Analysis and Sustainability Models

Chapter 6 discussed costs but didn't conduct detailed economic analysis across different deployment models. Future research should compare total cost of ownership for various approaches: commercial APIs vs. open-source local deployment, volunteer-based vs. automated geocoding, hybrid vs. fully automated workflows.

Cost analysis must go beyond nominal API pricing to include hidden expenses: technical staff time, infrastructure maintenance, volunteer recruitment and training, coordination overhead, and opportunity costs. Which models prove sustainable for which types of organizations? Can small grassroots groups afford any automated approach, or does automation primarily benefit well-funded international NGOs?

7.9 Temporal Dynamics and Address Evolution

This research used a static snapshot of 117 addresses. But urban geography changes constantly. New landmarks appear. Neighborhoods get renamed. Infrastructure develops. How do these temporal dynamics affect geocoding over time?

Longitudinal research could track how augmentation performance evolves as cities change. Do knowledge bases require continuous updating, or do they remain useful for months or years? When

major landmarks referenced in crisis reports get demolished or renamed, how quickly do systems adapt?

7.10 Integration with Other Crisis Informatics Challenges

Geocoding is one bottleneck in crisis mapping workflows. Future research should examine how LLM capabilities might address other challenges: verifying report credibility, extracting structured information from unstructured text, identifying duplicate reports, prioritizing urgent needs, and communicating with affected populations in their languages.

Perhaps the most powerful applications combine multiple capabilities. An LLM might simultaneously extract location information, identify the type of need (medical emergency vs. infrastructure damage), assess urgency based on language used, and flag reports requiring human verification. This holistic approach treats geocoding as one component in broader crisis informatics systems.

Each of these directions balances technical advancement with critical awareness. Better technology alone doesn't guarantee better outcomes. Research must continuously ask who benefits, who gets excluded, and how power shapes both problems and solutions. The goal isn't just improving metrics. It's building crisis response systems that genuinely serve the most vulnerable rather than optimizing existing inequalities.

7.11 Research Ethics and Positionality in Crisis Contexts

Chapter 6 addressed deployment ethics and governance structures. But the research process itself raises distinct ethical questions. How do we study crisis response technologies without treating disaster-affected communities as research subjects? What responsibilities do researchers bear when working with data generated during people's most vulnerable moments?

Future work must grapple with informed consent in crisis contexts. The 117 addresses used here came from a voluntary organization's operational records, anonymized and provided with institutional consent. But did the individuals who originally called for help consent to their location descriptions becoming research data? During active disasters, people report emergencies seeking rescue, not anticipating their words might inform academic studies. Even with anonymization, there's asymmetry. Researchers extract knowledge and career advancement from data generated during others' crises.

Participatory research approaches could address some of these tensions. What if affected communities weren't just data sources but co-researchers? Could longitudinal partnerships with local organizations create research relationships where communities shape questions, interpret findings, and benefit directly from outcomes? This requires patience and humility. It means accepting that communities might reject certain research directions as extractive or decide that improving geocoding isn't their priority.

Positionality matters too. I approached this research as a former volunteer, not a neutral outside observer. That insider perspective provided access and understanding. But it also shaped what questions seemed important and which solutions appeared viable. Future researchers should explicitly reflect on their own positions relative to the communities and crises they study. Technical objectivity doesn't erase the politics of who conducts research and whose knowledge counts as legitimate. Reflexivity about power dynamics in research relationships isn't optional. It's foundational to ethical practice.

8

Conclusion

This thesis began with a question that matters during disasters: where exactly are you? That question haunted me after my childhood experience in 2008, when I narrowly escaped floods that trapped others in the very places I had just visited. It drove my work with Robin Hood Army, where I spent five years helping the needy, coordinating relief, managing volunteers and learning that geocoding delays can mean the difference between rescue and tragedy. The elderly couple in Thanisandra, trapped with dwindling medication while volunteers spent hours trying to pinpoint their location, illustrated a pattern that repeats across disaster contexts. Their emergency should have taken minutes to map. It took two days.

Crisis mapping promised to democratize disaster response by enabling communities to report emergencies directly. Yet a persistent bottleneck emerged: translating informal location descriptions into coordinates that formal response systems understand. Haiti's 2010 earthquake demonstrated this challenge at devastating scale. Over 100,000 crisis reports poured into Ushahidi's platform. Only 3,584 were successfully mapped. Remote volunteers lacked the contextual knowledge to interpret descriptions of addresses. These weren't failures of effort or dedication. They exposed fundamental mismatches between how communities describe space and what geocoding systems require.

South Asian contexts intensify these challenges. Indian addressing evolved through centuries of landmark-based way-finding rather than administrative grids. Addresses conflate localities, landmarks, and social relationships without hierarchy. Bangalore exemplifies this complexity. My own house has five different address versions across official documents. Streets carry names in multiple languages and variations. New neighbourhoods emerge without appearing on any formal map. When addressing itself is fluid, geocoding during disasters becomes exponentially harder. Communities possessing the most accurate spatial intelligence often cannot translate it into formats response systems recognize.

Existing geocoding services fail dramatically on unstructured crisis addresses when compared to structured addresses in the global north (34). No systematic evaluation compared their performance in developing contexts where informal addressing dominates. Large Language Models

showed promise for bridging gaps between informal community knowledge and formal computational requirements through contextual reasoning. Yet their crisis geocoding application remained entirely unexplored, particularly for the low-resource, multilingual contexts where the need is most acute (35). Technical capability means little without deployment viability in resource-constrained settings where disasters disproportionately impact vulnerable populations.

Existing geocoding services fail dramatically on unstructured crisis addresses when compared to structured addresses in the global north (34). No systematic evaluation compared their performance in developing contexts where informal addressing dominates. Large Language Models showed promise for bridging gaps between informal community knowledge and formal computational requirements through contextual reasoning. One recent study applied LLMs to geolocalization, but in Western contexts with structured addressing conventions (35). Applying these techniques to informal, landmark-based spatial descriptions characteristic of South Asian crisis communications remained entirely unexplored. This research pioneers LLM-augmented geocoding evaluation for precisely these marginalized contexts: low-resource environments where communities communicate locations through colloquial landmarks, code-switched multilingual descriptions, and socially-embedded spatial relationships that defy formal standardization. Technical capability means little without deployment viability in resource-constrained settings where disasters disproportionately impact vulnerable populations.

This research addressed these interconnected gaps through three complementary research questions grounded in authentic Bangalore flood response data. I systematically compared five geocoding services (Google Maps, OLA Maps, OpenCage, Nominatim, Pelias) on 117 real crisis addresses, quantifying performance across operationally meaningful distance thresholds. I evaluated whether Large Language Model-based preprocessing could improve accuracy, testing eight augmentation techniques to identify effective strategies for crisis address standardization. I moved beyond technical performance to examine deployment feasibility, measuring costs, infrastructure requirements, and governance challenges for voluntary organizations with limited capacity. Together, these questions shifted focus from abstract technical potential toward grounded assessment of what actually works when internet fails, budgets constrain, and volunteers coordinate response using whatever tools survive.

Three findings challenge assumptions about automated crisis mapping. First, baseline performance reveals deep inequality. Google Maps and OLA Maps achieve moderate accuracy (63% and 46% of locations within 500m). OpenStreetMap-based services collapse spectacularly (2-5%). This is not random. Commercial services train and collect data on Indian addressing patterns that open-source alternatives lack. Training data embeds epistemological choices about whose spatial knowledge counts. Geography is not neutral. These services encode particular assumptions about legitimate ways to describe place.

Second, augmentation produces a paradox. LLMs improve OLA Maps substantially, reaching 68% precision with deterministic standardization. But they degrade Google Maps to 57%. Why the difference? Both services are proprietary, but the behavioural patterns suggest OLA's architecture

may expect structured input or that they have more context about India, with standardization and improvement potentially aligning messy addresses with its training distribution. Google appears to handle informal queries natively, possibly using contextual clues that standardization strips away. The lesson: no universal best practice exists. Effective augmentation depends on matching strategies to service architectures.

Third, technical metrics tell incomplete stories. Yes, OLA plus LLM augmentation outperforms Google's baseline by 5.3 percentage points. For crisis events processing thousands of reports, this means hundreds of additional accurate locations. Processing time drops from 5-10 minutes per address to seconds. During the floods in Bangalore, if there are about 1000 reports made to a voluntary organisation seeking help, improving precision from 62% to 68% means 60 additional locations accurately mapped. That is 60 more households rescue teams can reach without wasting critical time searching wrong neighbourhoods. These gains matter operationally.

But operational efficiency is not the only thing that matters. Automation shifts power in consequential ways. Communities gain natural language reporting but lose control over how their descriptions get interpreted. When the system misunderstands their neighbourhood consistently, they have no mechanism to contest or correct. NGOs reduce volunteer burnout but become dependent on AI infrastructure controlled by external entities. Pricing can spike. Services can shut down. Access can get restricted by policy changes or geopolitical tensions. LLM systems favour English and high-resource languages, systematically marginalizing communities reporting in Kannada, Tamil, or code-switched language.

Responsible deployment requires governance arrangements ensuring accountability, transparency, and genuine community participation. Technical architecture must balance automation efficiency with human oversight. Low-confidence geocodes should route to human review rather than appearing as certain. Communities need mechanisms to challenge systematic errors. Governance bodies cannot consist purely of technical experts making decisions about acceptable error rates. Communities bearing consequences of failures must have genuine voice in defining what counts as "good enough" accuracy. Multi-stakeholder bodies where community representatives hold genuine decision-making power, not just consultative roles. Transparency mechanisms that preserve original descriptions alongside standardized versions. Accountability structures enabling communities to flag systematic errors and get meaningful redress. Fairness monitoring that tracks performance differences across languages, neighbourhoods, and demographic groups.

The core tension: this technology improves efficiency within systems that already embed inequality. Digital divides exclude the most vulnerable from reporting at all. Training data encodes geographies where English place names and Western addressing receive more attention than Global South localities. Power imbalances determine whose needs get prioritized in response regardless of accurate geocoding. Calling LLM augmentation an equity intervention overstates the case. It removes one technical bottleneck. Most barriers to just crisis response lie in politics, institutions, history, and power relations that better geocoding cannot touch.

This research provides tools for informed deployment decisions: technical evidence showing what works in specific contexts, critical analysis revealing whose interests get served, and governance frameworks outlining what responsible implementation requires. The voluntary organization I worked with has expressed interest in piloting this approach during Bangalore's next monsoon season, recognizing both the operational improvements and the governance questions demanding ongoing attention. But deployment decisions belong to communities of practice seriously weighing capabilities against risks, not researcher prescription.

Ultimately, these questions return to real people facing real emergencies. The elderly couple in Thanisandra who waited days for help that should have arrived in hours. The Haiti survivors whose reports never reached rescue teams because geocoding could not keep pace. The Robin Hood Army volunteers I worked alongside, trying to coordinate flood response with whatever tools survived infrastructure collapse. Technical research validates that LLM-augmented geocoding improves accuracy by measurable margins for specific configurations in particular contexts. Whether those improvements translate to lives saved, faster response, and more equitable resource allocation depends entirely on how the technology gets governed, deployed, and held accountable to communities it claims to serve. That determination belongs to the communities and organizations doing the work. The answer to "where exactly are you?" matters most when someone's survival depends on getting it right. Technology can help us answer faster. It cannot substitute for the human judgment, local knowledge, and ethical commitment required to ensure that everyone, regardless of how they describe their location, receives the help they desperately need.

References

[1] The Hindu. **Hampi monuments remain flooded in Tungabhadra waters**. *The Hindu*, August 2022. Accessed: 2025-01-14. 1

[2] Nathan Morrow, Nancy Mock, Adam Papendieck, and Nicholas Kocmich. **Independent evaluation of the Ushahidi Haiti project**. *Development Information Systems International*, **8**(2011):111, 2011. 4, 9, 10

[3] Ida Norheim-Hagtun and Patrick Meier. **Crowdsourcing for crisis mapping in Haiti**. *Innovations: Technology/ Governance/ Globalization*, **5**(4):81, 2010. 4

[4] Matthew Zook, Mark Graham, Taylor Shelton, and Sean Gorman. **Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake**. *World Medical & Health Policy*, **2**(2):7–33, 2010. 4

[5] Vishal Srivastava, Priyam Tejaswin, Lucky Dhakad, Mohit Kumar, and Amar Dani. **A geocoding framework powered by delivery data**. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 568–577, 2020. 4, 6

[6] Bhavuk Singhal, Anshu Aditya, Lokesh Todwal, Shubham Jain, and Debashis Mukherjee. **GeoIndia: A Seq2Seq Geocoding Approach for Indian Addresses**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 395–407, 2024. 4, 5

[7] Brent Hecht and Monica Stephens. **A tale of cities: Urban biases in volunteered geographic information**. In *proceedings of the international AAAI conference on web and social media*, **8**, pages 197–205, 2014. 5, 8

[8] James C Scott. *Seeing like a state: How certain schemes to improve the human condition have failed*. yale university Press, 2020. 5, 8

[9] Andrea H Tapia, Kartikeya Bajpai, Bernard J Jansen, John Yen, and Lee Giles. **Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations.** In *ISCRAM*, 2011. 6

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. **On the dangers of stochastic parrots: Can language models be too big?** In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 6

[11] Leysia Palen and Kenneth M Anderson. **Crisis informatics—New data for extraordinary times**. *Science*, **353**(6296):224–225, 2016. 6

[12] Ory Okolloh. **Ushahidi, or'testimony': Web 2.0 tools for crowdsourcing crisis information**. *Participatory learning and action*, **59**(1):65–70, 2009. 9

[13] Jiri Panek, Lukas Marek, Vit Paszto, and Jaroslav Valuch. **The Crisis Map of the Czech Republic: the nationwide deployment of an Ushahidi application for disasters**. *Disasters*, **41**(4):649–671, 2017. 9, 11

[14] Ana Brandusescu and Renée E Sieber. **The spatial knowledge politics of crisis mapping for community development**. *GeoJournal*, **83**(3):509–524, 2018. 9, 10, 11

[15] Miren Gutierrez. **Maputopias: cartographies of communication, coordination and action—the cases of Ushahidi and InfoAmazonia**. *GeoJournal*, **84**:101–120, 2019. 10

[16] Robert Munro. **Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge**. In *Proceedings of the Workshop on Collaborative Translation: technology, crowdsourcing, and the translator perspective*, Denver, Colorado, USA, October 31 2010. Association for Machine Translation in the Americas. 10

[17] Khanh Ngo Duc, Tuong-Thuy Vu, and Yifang Ban. **Ushahidi and Sahana Eden open-source platforms to assist disaster relief: geospatial components and capabilities**. In *Geoinformation for Informed Decisions*, pages 163–174. Springer, 2014. 11

[18] Jessica Heinzelman and Carol Waters. *Crowdsourcing crisis information in disaster-affected Haiti*. JSTOR, 2010. 11

[19] Janet Marsden. **Stigmergic self-organization and the improvisation of Ushahidi**. *Cognitive Systems Research*, **21**:52–64, 2013. 11

[20] Angela Crandall and Rhoda Omenya. **Uchaguzi Kenya**. *Biographies 3*, page 69, 2015. 11

[21] NAVO Kaushalye and S Koswatte. **Crowdsourced Data Relevance Analysis for Crowd-assisted Flood Disaster Management**. *Journal of Geospatial Surveying*, **1**(1), 2021. 11

[22] S Koswatte, K McDougall, and X Liu. **Crowd-assisted flood disaster management**. In *Application of Remote Sensing and GIS in Natural Resources and Built Infrastructure Management*, pages 39–55. Springer, 2023. 11

[23] Team Ola. **Navigating India - The Journey of Ola Maps**. https://tech.olakrutrim.com/navigating-india-the-journey-of-ola-maps/, July 2024. Accessed: 2025-11-03. 52

[24] Google Cloud. **Google Maps Platform Pricing**. https://mapsplatform.google.com/pricing/, 2025. Accessed: 2025-01-09. 61, 62

[25] Ola Krutrim. **Pricing - Ola Maps API Plans for Businesses & Developers**. https://maps.olakrutrim.com/pricing, 2025. Accessed: 2025-01-09. 61, 62

[26] Amazon Web Services. **Amazon Bedrock Pricing**. https://aws.amazon.com/bedrock/pricing/, 2025. Accessed: 2025-01-09. 61, 62

[27] Chelsea Barabas. **Care as (re) capture: Data colonialism and race during times of crisis**. *New Media & Society*, **26**(12):7351–7370, 2024. 63

[28] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. **Bias and fairness in large language models: A survey**. *Computational Linguistics*, **50**(3):1097–1179, 2024. 64

[29] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. **A survey on multilingual large language models: Corpora, alignment, and bias**. *Frontiers of Computer Science*, **19**(11):1911362, 2025. 64

[30] Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. **Diversity and language technology: how language modeling bias causes epistemic injustice**. *Ethics and Information Technology*, **26**(1):8, 2024. 64

[31] Yuvraj Gupta, Zhewei Liu, and Ali Mostafavi. **Digital Divide in Disasters: Investigating Spatial and Socioeconomic Disparities in Internet Service Disruptions During Extreme Weather Events**. *arXiv preprint arXiv:2312.08640*, 2023. 64

[32] Jennifer S Dargin, Chao Fan, and Ali Mostafavi. **Vulnerable populations and social media use in disasters: Uncovering the digital divide in three major US hurricanes**. *International Journal of Disaster Risk Reduction*, **54**:102043, 2021. 64

[33] Evgeny Morozov. **To save everything, click here: the folly of technological solutionism**. *J. Inf. Policy*, **4**(2014):173–175, 2014. 65

[34] Błażej Ciepłuch, Ricky Jacob, Peter Mooney, and Adam C Winstanley. **Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps**. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, page 337. University of Leicester, 2010. 79, 80

[35] Wenping Yin, Yong Xue, Ziqi Liu, Hao Li, and Martin Werner. **LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of Hurricane Harvey**. *International Journal of Applied Earth Observation and Geoinformation*, **137**:104423, 2025. 80

9

Appendix

9.1 LLM Augmentation Prompt Templates

This appendix provides complete prompt templates for all eight LLM-based address augmentation techniques evaluated in Section 3.3. All prompts target GPT-OSS-20B-1 via AWS Bedrock with temperature 0.2 (T1-T7) or temperature 0.0 (T8). Prompts mandate JSON-only output to ensure parseable responses.

9.1.1 T1: Zero-Shot Baseline

```
Standardize this Bangalore address for geocoding.

Address: {address}

Apply these steps:
1. Split concatenated tokens (e.g., "NagsenNagar" -> "Nagsen Nagar")
2. Correct spelling errors
3. Identify building numbers
4. Remove delivery instructions and phone numbers
5. Keep only location components

Output format: [Building], [Street], [Sub-locality], [Locality],
               City, State, PIN, Country

Output ONLY JSON (no explanation):
{"address": "your cleaned address here"}
```

Figure 9.1: T1: Zero-Shot Baseline Prompt Template

Technique characteristics:

- Prompting strategy: Zero-shot with explicit instructions

- No examples provided
- Five-step standardization protocol
- Structured output format specification
- Temperature: 0.2

9.1.2 T2: Few-Shot Learning

```
Standardize using these examples:

Input: "N/N 2ndstreet govindaswmy layout PKDNagar"
Output: {"address": "N/N, 2nd Street, Govindaswamy Layout,
         PKD Nagar, Bangalore, Karnataka, India"}

Input: "Door No. N-N-N, Ground Floor, East Point Colony
       (Mobile: 11111-11111)"
Output: {"address": "Door No. N-N-N, Ground Floor,
         East Point Colony, Bangalore, Karnataka, India"}

Input: "HSR 6th sector near BDA complex"
Output: {"address": "6th Sector, BDA Complex, HSR Layout,
         Bangalore, Karnataka, India"}

Now standardize: {address}

Output ONLY JSON:
{"address": "your cleaned address here"}
```

Figure 9.2: T2: Few-Shot Learning Prompt Template

Technique characteristics:

- Prompting strategy: Few-shot learning with 3 examples
- Demonstrates: concatenated token handling, door number formatting, landmark-based descriptions
- Examples show pattern diversity (residential, commercial, landmark-based)
- No explicit instructions - model learns implicitly from examples
- Temperature: 0.2

9.1.3 T3: Context-Augmented RAG

```
BANGALORE ADDRESSING KNOWLEDGE BASE

=== REGIONAL TERMINOLOGY ===
Street Types: Road, Street, Avenue, Main Road, Cross, Circle
Landmarks: Complex, Layout, Nagar, Colony, Extension, Stage
Building Types: Building, Apartment, Residency, Tower, Block
Administrative: Ward, Zone, Taluk, Hobli, Division

=== AREA ALIASES ===
Koramangala: Koramangla, Kormangala, KMG
HSR Layout: HSR, Sector 1-7, BDA Complex Area
Indiranagar: Indira Nagar, I Nagar, Defense Colony Area
Whitefield: White Field, ITPL Area, Hope Farm
Jayanagar: Jaya Nagar, J Nagar, Blocks 1-9

=== COMMON MISSPELLINGS ===
Bangalore: Bangaluru, Banglore, Bengaluru
Koramangala: Koramangla, Kormangla
Indiranagar: Indranagar, Indira Nagar
Marathahalli: Maratahalli, Maratha Halli
Yelahanka: Yellahanka, Yalahanka

Use this knowledge to standardize: {address}

Apply: correct misspellings, expand aliases, identify chunks,
       reconstruct hierarchy.

Output ONLY JSON:
{"address": "your cleaned address here"}
```

Figure 9.3: T3: Context-Augmented RAG Prompt Template

Technique characteristics:

- Prompting strategy: Retrieval-augmented generation (RAG)
- Knowledge base components: regional terminology, area aliases, common misspellings
- Bangalore-specific context (locality hierarchies, landmark patterns)
- Enables correction of phonetic variants and informal names
- Temperature: 0.2

Note: The knowledge base shown represents a condensed version. The full knowledge base contains 50+ area aliases, 100+ misspelling patterns, and comprehensive terminating token categories extracted from the training dataset.

9.1.4 T4: Combined System Prompting

```
You are an address standardization expert for Bangalore, Karnataka.
You understand Indian address patterns, regional terminology, and
common errors.

[KNOWLEDGE BASE CONTENT - Same as T3 above]

Address: {address}

Protocol:
1. De-concatenate joined tokens
2. Spell-check using knowledge base
3. Identify chunks using terminating tokens
4. Reconstruct: Building -> Street -> Sub-locality -> Locality
                -> City -> State
5. Add missing components (Bangalore, Karnataka, India)

Output ONLY JSON (no explanation):
{"address": "your cleaned address here"}
```

Figure 9.4: T4: Combined System Prompting Template

Technique characteristics:

- Prompting strategy: System prompting + RAG context + explicit protocol
- Assigns expert role (address standardization specialist)
- Combines knowledge base (T3) with structured instructions (T1)
- Five-step protocol emphasizes hierarchy reconstruction
- Temperature: 0.2

9.1.5 T5: Chain-of-Thought Reasoning

```
Standardize this Bangalore address using BRIEF reasoning:

Address: {address}

BRIEF STEPS:
1. Tokenize & split concatenated tokens
2. Correct misspellings
3. Identify chunks: building, street, sub-locality, locality
4. Remove extraneous info (phones, instructions)
5. Reconstruct in hierarchy order

Keep reasoning concise (1-2 sentences per step).

Final output in JSON:
{"address": "your cleaned address here"}
```

Figure 9.5: T5: Chain-of-Thought Reasoning Prompt Template

Technique characteristics:

- Prompting strategy: Chain-of-thought reasoning
- Requests explicit step-by-step reasoning (1-2 sentences per step)
- Five reasoning steps mirror T1 protocol
- Balances reasoning visibility with output conciseness
- Temperature: 0.2

9.1.6 T6: Iterative Refinement

First Iteration (iteration = 1):

```
[Uses T4 Combined technique prompt as initial pass]
```

Figure 9.6: T6: Iterative Refinement - First Pass

Subsequent Iterations (iteration > 1):

```
ITERATION {iteration}:
Original: {address}
Previous: {previous_result}
Error: {geocoding_error}m

[If error > 500m:]
Geocoding error > 500m. Add more specific components or try
alternative aliases.

Apply refinements. Output ONLY JSON:
{"address": "your cleaned address here"}
```

Figure 9.7: T6: Iterative Refinement - Subsequent Iterations with Feedback

Technique characteristics:

- Prompting strategy: Iterative refinement with geocoding feedback
- First iteration: T4 Combined technique (baseline augmentation)
- Subsequent iterations: Receive previous result and positional error
- Conditional feedback: If error $>$ 500m, suggests adding components or alternative names
- Enables error-driven refinement loop
- Temperature: 0.2

Implementation note: In evaluation, only single-iteration results (T4 baseline) are reported due to computational constraints. Multi-iteration refinement remains available for future exploration.

9.1.7 T7: Role-Based Expert Persona

```
You are Dr. Priya Sharma, a GIS specialist with 15 years experience
in Bangalore addresses.

Expertise: Bangalore addressing conventions, regional terminology,
linguistic variations (Hindi/Kannada/English), urban planning layouts.

Mission: Emergency response system needs precise location. Lives
depend on accuracy.

Address: {address}

Tasks: Correct phonetic errors, expand abbreviations, remove
non-location info, ensure logical hierarchy.

Output ONLY JSON (no explanation):
{"address": "your cleaned address here"}
```

Figure 9.8: T7: Role-Based Expert Persona Prompt Template

Technique characteristics:

- Prompting strategy: Role-based prompting with expert persona
- Persona: Dr. Priya Sharma (fictitious GIS specialist)
- Expertise dimensions: regional conventions, linguistic variations, urban planning
- Mission framing: Emergency response context emphasizing accuracy importance
- Leverages role-playing capability of large language models
- Temperature: 0.2

9.1.8 T8: Deterministic Rule-Based Protocol

```
DETERMINISTIC PROTOCOL - Follow rules EXACTLY:

1. Normalize: lowercase, remove special chars (keep , / : #),
              collapse spaces
2. Process tokens: split concatenated, correct misspellings,
                   expand abbreviations
3. Extract: building number, road, locality indicators
4. Filter: remove phone numbers, delivery instructions
5. Assemble: [Building], [Street], [Sub-locality], [Locality],
             Bangalore, Karnataka, India

Address: {address}

NO explanation. Output ONLY JSON:
{"address": "your cleaned address here"}
```

Figure 9.9: T8: Deterministic Rule-Based Protocol Template

Technique characteristics:

- Prompting strategy: Deterministic rule-based processing
- Temperature: **0.0** (only technique using zero temperature)
- Explicit "EXACTLY" emphasis for strict adherence
- Five-step deterministic protocol
- Designed for maximum reproducibility and consistency
- Emphasizes rule-following over creative interpretation

9.1.9 Implementation Details

9.1.9.1 Model Configuration

All techniques use GPT-OSS-20B-1 via AWS Bedrock with the following parameters:

Table 9.1: LLM Configuration Parameters

Parameter	Value
Model	GPT-OSS-20B-1
API	AWS Bedrock
Temperature (T1-T7)	0.2
Temperature (T8)	0.0
Max tokens	512
Top-p	1.0
Frequency penalty	0.0
Presence penalty	0.0
Stop sequences	None

9.1.9.2 Output Format

All techniques enforce JSON-only output with the following schema:

```
{
  "address": "standardized address string"
}
```

Figure 9.10: Expected JSON Output Schema

Responses failing to parse as valid JSON are treated as augmentation failures. The `address` field contains the complete standardized address string used for geocoding service queries.

9.1.9.3 Prompt Variable Substitution

The placeholder `{address}` in all templates is replaced with the original unstructured address string from the dataset. No preprocessing occurs before prompt construction, original addresses are inserted verbatim, preserving concatenation errors, misspellings, and informal descriptions.

9.1.9.4 Knowledge Base Source

The knowledge base for T3 and T4 is constructed from:

- Bangalore OpenStreetMap (OSM) data exports (administrative boundaries, landmark names)
- Manual curation of common misspellings identified in training data
- Regional terminology extracted from official Karnataka addressing guidelines
- Area aliases collected from local community forums and real estate listings

The complete knowledge base JSON file (`bangalore_addressing_kb.json`) contains:

- 8 terminating token categories (street types, landmarks, building types, administrative units)

- 50+ area canonical names with aliases
- 100+ misspelling patterns
- Token frequency distributions from 1000+ Bangalore addresses

9.1.9.5 Reproducibility

To reproduce these experiments:

1. Use identical prompt templates as specified above
2. Configure GPT-OSS-20B-1 with parameters from Table A.1
3. Ensure knowledge base content matches T3/T4 specifications
4. Apply temperature 0.2 for T1-T7, temperature 0.0 for T8
5. Parse JSON responses; treat invalid JSON as augmentation failure
6. Submit standardized addresses to geocoding services using identical API versions

All code, prompts, and knowledge base files are available in the supplementary materials repository.

9.1.10 Design Rationale

The eight techniques represent a progression through modern LLM prompting strategies:

- **T1 (Zero-Shot)**: Baseline prompting with explicit instructions - establishes minimum augmentation capability
- **T2 (Few-Shot)**: Tests in-context learning - assesses whether examples improve standardization
- **T3 (RAG)**: Evaluates knowledge augmentation - determines if domain-specific context enhances accuracy
- **T4 (Combined)**: Combines system prompting, knowledge, and protocol - represents "best practices" synthesis
- **T5 (Chain-of-Thought)**: Investigates reasoning visibility - examines if explicit reasoning improves results
- **T6 (Iterative)**: Explores feedback-driven refinement - tests error-corrective iteration capability
- **T7 (Role-Based)**: Applies persona-based prompting - leverages role-playing for task-specific expertise

- **T8 (Deterministic)**: Establishes reproducibility baseline - maximizes consistency at temperature 0

This design variety enables comprehensive evaluation across the spectrum of LLM augmentation approaches, from minimal intervention (T1, T8) through knowledge-enhanced processing (T3, T4) to sophisticated reasoning strategies (T5, T6, T7).