Geographic Inference with Large Language Models: Identification of POIs and Urban Features from Addresses and Coordinates

Gabriel A. C. Silva[1], **Salatiel D. Silva**[1], **Claudio E. C. Campelo**[1]

[1] Systems and Computing Department
Federal University of Campina Grande (UFCG) Campina Grande – PB – Brazil

gabriel.alves@copin.ufcg.edu.br,
{salatiel.dantas,campelo}@computacao.ufcg.edu.br

Abstract. *This paper evaluates the ability of Large Language Models (LLMs) to infer geographic features and identify points of interest (POIs) from addresses and coordinates. Nine cities of different urban sizes, distributed across three continents, were analyzed, with five addresses per city in scenarios such as urban center, residential area, green area, proximity to water and mixed area. The predictions were validated with external geospatial databases. The results indicate greater stability in feature inference than in POI identification, whose performance varies according to country and urban size; Gemini 3.1 Flash Lite leads in large cities, GPT-4.1 leads in medium-sized cities, and Gemini 3 Flash Preview leads in small cities.*

1. Introduction

Large Language Models (LLMs) have shown strong performance in text generation, comprehension, and reasoning tasks. In recent years, these models have also been investigated in problems that require knowledge about the physical world, including tasks involving location [Manvi et al. 2024], place description and urban context interpretation [Feng et al. 2025, Yan and Lee 2024]. This interest stems from the fact that a significant portion of the knowledge available in Web texts contains explicit or implicit spatial references, which may allow LLMs to capture geographic, socioeconomic, and urban patterns during training [Manvi et al. 2024].

Despite this potential, geographic inference with LLMs still presents important challenges. Unlike purely textual tasks, geospatial problems depend on metric relations, direction, proximity, urban hierarchy, and the updating of local information, such as the existence of points of interest (POIs). Recent work shows that general-purpose models can incorporate relevant spatial knowledge, but also reveals limitations when they are evaluated at finer urban scales or in regions underrepresented in training data [Feng et al. 2025, Moayeri et al. 2024]. In addition, approaches that combine LLMs with structured databases, such as OpenStreetMap (OSM), indicate that the parametric knowledge of models may be insufficient when used in isolation to solve specific geographic tasks [Yan and Lee 2024].

Although previous studies evaluate LLMs in geospatial tasks, these evaluations generally rely on specific settings, such as regional benchmarks, models adapted to the urban domain, or approaches supported by structured geographic databases

[Feng et al. 2025, Moayeri et al. 2024, Yan and Lee 2024, He et al. 2025]. In this context, it remains underexplored to what extent general-purpose LLMs can infer urban features and identify POIs using only textual addresses and geographic coordinates, without consulting external sources during response generation.

Given these limitations, this paper evaluates whether general-purpose LLMs can produce useful geographic inferences when they receive only basic location information. The analysis considers two tasks: inferring geographic features and characteristics of the surrounding area from an address and its coordinates; and listing nearby POIs, including type, direction, and estimated distance. To this end, we evaluate scenarios in cities from different countries and continents, considering contexts such as urban centers, residential areas, green areas, areas near water, and mixed-use areas. The responses are subsequently verified against external geospatial sources, enabling the evaluation of both the presence of inferred features and the existence of generated POIs.

2. Related Work

LLMs have been widely investigated in geospatial tasks, spanning geospatial reasoning, urban context interpretation, geographic bias, and location-based representation learning. Current literature, however, predominantly focuses either on augmenting models with external geospatial databases or on assessing spatial awareness at a regional scale, whereas this study evaluates standalone general-purpose LLMs at the neighborhood level using only addresses and coordinates as input.

General-purpose models frequently struggle with detailed geographic tasks. The *CityGPT* framework proposed by [Feng et al. 2025] integrates urban data and specific instructions to improve the spatial cognition of LLMs, including tasks such as navigation and city-level spatial relation inference [Feng et al. 2025]. Unlike CityGPT, the present work evaluates models under strict parametric-only conditions, with no external data provided during generation, in order to isolate what LLMs can infer from location information alone.

In the context of integrating natural language and geographic data, GeoReasoner [Yan and Lee 2024] combines linguistic information with structured geospatial databases, such as OSM, to improve tasks such as geographic entity recognition and disambiguation [Yan and Lee 2024]. In contrast, the present work uses OSM-based data only for post-hoc validation, deliberately preventing models from accessing any external source during response generation.

Prior work has also directly probed the geospatial knowledge encoded in LLMs. [Bhandari et al. 2023] evaluate coordinate prediction, sensitivity to spatial prepositions, and spatial reasoning via multidimensional scaling, showing that larger models perform better but still exhibit limitations in localization and geographic reasoning. Beyond task performance, [Liu et al. 2024] examine geographic diversity using a natural-language geo-guessing experiment with GPT-4 and structured knowledge base abstracts, finding insufficient knowledge for several feature types and variation across scale, region, and model variant. While these studies operate at global or regional scales, the present work evaluates performance at the neighborhood level across predefined urban scenario types, enabling a more granular diagnosis of where LLM spatial knowledge degrades.

Another line of research focuses on generating spatial representations from

LLMs. In this context, LLMGeovec [He et al. 2025] uses textual descriptions derived from geographic coordinates to produce embeddings that capture spatial and semantic characteristics, applied to tasks such as spatio-temporal prediction and geographic analysis. Rather than using LLM outputs as feature representations, the present work directly evaluates their factual correctness against external geospatial ground truth, assessing the accuracy of inferred features and identified POIs.

On the other hand, recent studies also highlight important limitations of these models. The WorldBench benchmark [Moayeri et al. 2024] shows disparities in LLM performance across different geographic regions, with higher error rates in countries underrepresented in training data [Moayeri et al. 2024]. The present work extends this concern by explicitly controlling for urban size and linking geographic bias to specific task types, namely geographic feature inference and POI identification, rather than to factual recall alone.

Taken together, these studies motivate an evaluation of whether broad geospatial knowledge in LLMs translates into reliable neighborhood-level judgments, especially for feature inference and POI verification under controlled prompting.

3. Methodology

This work adopts an experimental approach to evaluate LLMs in geographic feature inference and POI identification from location information. We compare models from Google, OpenAI, Anthropic, and DeepSeek, representing distinct model families. In all experiments, the models were executed without access to external sources during generation and, when allowed by the API, with temperature set to zero. Search, *grounding*, external tools, and Web browsing resources were disabled; external geospatial databases were used only for response validation.

3.1. City and Scenario Selection

Nine cities distributed across North America, South America, and Europe were selected, considering different levels of urban complexity: New York, Vancouver, and Aspen in North America; São Paulo, Campina Grande, and Tibaú do Sul in South America; and London, Amsterdam, and Bath in Europe. These cities were chosen based on data availability in the validation databases, ensuring consistent verification and minimizing discrepancies from missing data. This balances geographic and urban diversity with external validation feasibility. Thus, the selection seeks to favor a fairer comparison across models, maintaining geographic and urban diversity without compromising the feasibility of external validation.

In addition to geographic distribution, the selection considered different urban sizes, as city size influences urban density, POI diversity, and the complexity of surrounding features; larger cities also tend to have more textual and geospatial data that may be present in model training. To this end, population size was defined using an operational classification inspired by the typology of the Organisation for Economic Co-operation and Development (OECD), an international organization that provides comparable socioeconomic indicators across countries for urban areas by population size. Based on the OECD indicator [OECD 2026], metropolitan population was considered when available and municipal population otherwise. Thus, cities were grouped into small,

medium, and large categories as a criterion of sample diversity in density and spatial organization. For each city, five scenarios were defined: urban center, residential area, green area, area near water, and mixed-use area, ensuring diversity of geographic contexts and totaling 45 experimental scenarios (nine cities $\times$ five scenarios) per model.

3.2. Models and Prompts

Models from different families and organizations were evaluated: Gemini 3.1 Flash Lite and Gemini 3 Flash Preview, from Google; DeepSeek V4-Pro and DeepSeek V4-Flash, from DeepSeek; Claude Sonnet 4.6, from Anthropic; and GPT-4.1, from OpenAI. All models received the same set of inputs and were submitted to the same prompts, enabling their behavior to be compared under a common experimental configuration. The prompts, scripts, and aggregated data are available in a public repository[1].

In the first experiment, the models receive a prompt to generate a structured description of the physical environment and infer geographic features within a 200-meter radius of the reference location. This radius focuses the analysis on the immediate surroundings and reduces the inclusion of distant elements that might not characterize the local scenario. The prompt considers predefined categories, such as urban density, green areas, and road access, and imposes formatting and content restrictions to standardize and compare the responses.

In the second experiment, the prompt asks the model to list exactly ten plausible POIs within up to two kilometers of walking distance from the reference location, reporting name, type, estimated distance, and cardinal or ordinal direction. The types are restricted to predefined categories, prioritizing real and specific places, such as urban landmarks, known streets, transit stations, parks, commercial areas, and public places. These rules reduce generic, distant, or nonexistent responses, such as *museum*, *central square*, or *subway station* without a specific name, outside the defined radius, or without a verifiable relation to the analyzed point. Walking distance brings the task closer to real urban travel, avoiding limitations of Euclidean distance, which ignores street layout, physical barriers, blocks, rivers, parks, or missing direct connections, as well as vehicle-based estimates, which depend on traffic, road direction, circulation restrictions, and local conditions.

3.3. Results Analysis

Responses are validated against external geospatial references. Geographic features are verified via the Overpass API (OSM data), with exact tag-matching schemas and query definitions hosted in our public repository. For POIs, candidates are fetched via the Google Places API within a ten km radius, but successful matches require confirmation within a strict 2 km walking distance, directly aligning with the prompt constraint. Given that models are restricted to exactly ten generated items, we evaluate a response-level, slot-based F1-score to measure POI identification performance rather than macro-scale spatial recall. Under this framework, feature predictions are true positives (TP) if confirmed by OSM, false positives (FP) if unconfirmed, and false negatives (FN) if omitted. A predicted POI is a TP if verified within the two km walking-distance limit, an FP if unverified or located only between two and ten km, and an FN if a requested response slot remains unverified, totaling 50 expected slots per city (ten per scenario).

[1]Repository available at: `https://github.com/gabrizl/llm-geographic-inference`.

4. Experiments and Results

This section presents country-level aggregated results for the F1-score of geographic features and POIs. Next, POI metrics are analyzed by city size, considering the identification F1-score. The interpretations are organized as aggregations by country and by urban size, highlighting comparative patterns across models and urban contexts.

4.1. Geographic Feature Inference

In geographic feature inference, the models show moderate to high performance (Table 1). Considering the average F1-scores by country, Gemini 3 Flash Preview achieves the best overall result (0.701), followed by Gemini 3.1 Flash Lite (0.668) and Sonnet 4.6 (0.655). The distribution of the best results by country is also concentrated in these three models: Gemini 3 Flash Preview leads in Brazil and the United States and ties with Gemini 3.1 Flash Lite in Canada, while Sonnet 4.6 obtains the highest F1-scores in the Netherlands and England.

Table 1. F1-score of geographic feature inference across the evaluated countries. The best results by country are highlighted in bold.

Country	Gemini 3 Flash Preview	Sonnet 4.6	Gemini 3.1 Flash Lite	DeepSeek V4-Flash	GPT-4.1	DeepSeek V4-Pro
Brazil	**0.644**	0.559	0.554	0.551	0.548	0.448
Canada	**0.791**	0.625	**0.791**	0.600	0.700	0.581
Netherlands	0.588	**0.621**	0.612	0.526	0.524	0.343
England	0.767	**0.769**	0.725	0.554	0.676	0.600
United States	**0.716**	0.702	0.658	0.582	0.600	0.667

The spread across models is more pronounced in the Netherlands (range of 0.278 between the highest and lowest F1-score) and in Brazil (0.196), while the United States shows the lowest dispersion (0.134). Sonnet 4.6 is the only model outside the Google family to lead in any country, obtaining the highest F1-scores in the Netherlands (0.621) and England (0.769), showing that the distribution of best results is not uniform across model families.

4.2. POI Identification

POI identification displays higher volatility across regions and models (Table 2). The slot-based F1-score balances precision and recall over the requested response items, penalizing unverified or omitted slots. Regionally, performance trends fluctuate: GPT-4.1 achieves the highest F1-scores in Brazil (0.300) and Canada (0.560), while DeepSeek V4-Pro leads in the United States (0.590). Additionally, Gemini 3 Flash Preview tops the sample in England (0.613) and Gemini 3.1 Flash Lite leads in the Netherlands (0.420).

Table 2. F1-score of POI identification by country. The best results by country are highlighted in bold.

Country	Gemini 3 Flash Preview	Sonnet 4.6	Gemini 3.1 Flash Lite	DeepSeek V4-Flash	GPT-4.1	DeepSeek V4-Pro
Brazil	0.245	0.193	0.273	0.267	**0.300**	0.193
Canada	0.505	0.400	0.480	0.420	**0.560**	0.540
United States	0.429	0.390	0.490	0.540	0.480	**0.590**
England	**0.613**	0.474	0.540	0.580	0.500	0.580
Netherlands	0.265	0.300	**0.420**	0.340	0.400	0.400

4.3. Analysis by City Size

Table 3 shows that POI identification varies across urban dimensions, revealing shifts in model performance hierarchies.

Table 3. F1-score of POI identification by city size. The best results by size are highlighted in bold.

Size	Gemini 3 Flash Preview	Sonnet 4.6	Gemini 3.1 Flash Lite	DeepSeek V4-Flash	GPT-4.1	DeepSeek V4-Pro
Large	0.494	0.429	**0.567**	0.560	0.527	0.520
Medium	0.324	0.326	0.400	0.373	**0.433**	0.413
Small	**0.400**	0.253	0.293	0.333	0.313	0.353

Gemini 3.1 Flash Lite leads in large cities (0.567), GPT-4.1 in medium contexts (0.433), and Gemini 3 Flash Preview in small cities (0.400). Rather than showing a linear progression, this pattern suggests that POI identification is sensitive to localized data distributions across scales. Thus, standalone LLMs may support exploratory urban analyses, but their uneven performance on granular assets reinforces the need for cross-referencing parametric outputs with independent geospatial databases.

5. Final Considerations

The results suggest greater stability in the inference of general urban features than in POI identification. For POIs, the best performances vary according to country and urban size: Gemini 3.1 Flash Lite leads in large cities, GPT-4.1 leads in medium-sized cities, and Gemini 3 Flash Preview leads in small cities. This variation suggests the combined influence of validation database coverage, urban density, and the availability of local knowledge in the models' internal data.

As future work, we intend to expand the set of models, cities, and countries evaluated, including English-speaking African countries, in order to investigate possible biases associated with language in the models' training data. We also intend to conduct a qualitative analysis of the errors generated by the models, considering both geographic feature inference and POI identification, including inaccuracies in POI names and estimated distances. The responses may also be compared with multiple geospatial databases and explored in geographic recommendation systems, provided that they are combined with external validation, updating, and spatial calculation procedures.

References

Bhandari, P., Anastasopoulos, A., and Pfoser, D. (2023). Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, New York, NY, USA. Association for Computing Machinery.

Feng, J., Liu, T., Du, Y., Guo, S., Lin, Y., and Li, Y. (2025). Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 591–602, New York, NY, USA. Association for Computing Machinery.

He, J., Nie, T., and Ma, W. (2025). Geolocation representation from large language models are generic enhancers for spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17094–17104.

Liu, Z., Janowicz, K., Currier, K., and Shi, M. (2024). Measuring geographic diversity of foundation models with a natural language–based geo-guessing experiment on gpt-4.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., and Ermon, S. (2024). Geollm: Extracting geospatial knowledge from large language models. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y., editors, *International Conference on Learning Representations*, volume 2024, pages 38791–38807.

Moayeri, M., Tabassi, E., and Feizi, S. (2024). Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1211–1228, New York, NY, USA. Association for Computing Machinery.

OECD (2026). Urban population by city size. `https://www.oecd.org/en/data/indicators/urban-population-by-city-size.html`. Acesso em: 16 Abril 2026.

Yan, Y. and Lee, J. (2024). Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4163–4167, New York, NY, USA. Association for Computing Machinery.