

---

# Use of Pretrained Language Models for Geographic Information Retrieval

Accepted Abstract Poster, GIScience 2025

In: Proc. Thirteenth International Conference on Geographic Information Science (GIScience 2025)

August 26–29, 2025  
Ōtautahi | Christchurch  
Aotearoa | New Zealand

Alexis Horde Vo<sup>1,2,\*</sup>, Matt Duckham<sup>1</sup>, Estrid He<sup>2</sup>, Nayomi Geethanjali Ranamuka<sup>1,2</sup>, Rafe Benli<sup>3</sup>

1. Geographic Knowledge Lab, School of Science, RMIT University, Australia
2. School of Computing Technologies, RMIT University, Australia
3. Geographic Names Victoria, Victoria State Government, Australia

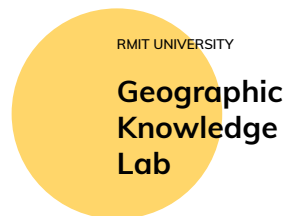
\*alexis.horde.vo@student.rmit.edu.au

## ABSTRACT

Pre-trained language models (PLMs) are developed through extensive observations of co-occurrent linguistic patterns. By solving many natural language processing (NLP) tasks, PLMs have demonstrated remarkable generalization capabilities, with extended usages in Geographic Information Retrieval (GIR). On a specific case, we study what specificities of GIR are still uncovered by PLMs. We found that PLMs are biased towards a better semantic understanding to the detriment of spatial matching. Indeed, the set of retrieved answers are often spatially disconnected from the spatial context of a query.

The specific study case used in this work is the identification of place name origins in Melbourne (Australia). We compared a naïve retrieval-augmented generation (RAG) architecture—without fine-tuning PLMs—with a spatially oriented RAG one, so-called “geoRAG”. Both RAG architectures use two PLMs: ColBERTv2 and Llama2-13B. Contrary to a naïve RAG, geoRAG aligns a query spatially and semantically with candidates from the knowledge graph DBpedia. Indeed, geoRAG is structured as an ensemble model: for each candidate-answer, semantic matching and spatial matching are distinctly treated and scored by a finetuned ColBERTv2 on Geonames. Both resulting scores are combined in a late fusion to extract the top-*k* candidate-answers. As a final step, Llama2-13B interprets the results and generates the final answer.

When specifically evaluating spatial matching, the naïve RAG does not consider space as a hard constraint. In contrast, geoRAG tackles the problem by underlying space as a specific modality: an increase of 16% in normalized discounted cumulative gain (nDCG) is observed for the spatial



component. This high gain reveals that the geographic structure of the world is not intrinsically indexed in PLMs in traditional information retrieval: answers might be spatially heterogeneous and incompatible. In conclusion, one main characteristic of GIR systems is to find what is spatially close as a first objective, where PLMs still struggle.

#### **ACKNOWLEDGMENTS**

This research was undertaken with the assistance of computing resources from RACE (RMIT Advanced Cloud Ecosystem).

The authors would like to acknowledge Nenad Radosevic for his insightful help at the Geographic Knowledge Lab (RMIT University).