MASTS Visiting Fellowship — Project Report

Fellow: Dr. Cai Wu, The Hong Kong University of Science and Technology (Guangzhou)
Host(s): Dr. Mingshu Wang, University of Glasgow

Coastal Urban Activities & Marine Ecosystems: building an evidence base with literature mining, LLM extraction, and Scottish field engagement

Project aims

This project seeks to advance understanding of the interactions between coastal and marine outcomes and urban pressures through a combination of data synthesis, methodological innovation, and collaborative engagement.

First, we will build a comprehensive database of current research by systematically linking coastal and marine outcomes to urban drivers, drawing from Scopus-indexed literature. Second, we aim to prototype a large language model (LLM)–assisted analytical pipeline capable of inferring study areas and thematic focuses directly from publication titles and abstracts, thereby accelerating knowledge mapping and synthesis.

Third, the project will strengthen Scotland-based collaborations through targeted visits, joint supervision of students, thematic workshops, and reconnaissance of key coastline sites to ground emerging research questions in real-world contexts. Finally, we will conduct a foresight and equity analysis to identify future hotspots of human activity along global coastlines and to assess whether scholarly attention aligns with projected population distributions and exposure risks.

Activities & Progress

A) Data collection from Scopus

The initial phase of the project, focused on data collection from Scopus, has been successfully completed. The literature selection followed a clearly defined set of inclusion and exclusion criteria to ensure relevance and analytical consistency. Studies were included if they (i) addressed coastal or estuarine settings; (ii) examined explicit urban drivers or proxies such as ports, wastewater, shipping, or urbanization; (iii) were based on empirical sites or contained mappable geographic contexts; and (iv) were published from 2000 onwards.

Publications were excluded if they dealt solely with offshore or pelagic environments lacking urban linkages, were purely conceptual without site-specific data, or represented duplicate records.

The resulting dataset was characterized using Scopus “Analyze” summary tools, which provide an overview of the top disciplinary categories and their counts. Percentages reported in this analysis refer to proportions within the selected top categories, rather than the entire global corpus.

By Country

Rank	Country	Share of listed set
1	United States	15.85%
2	China	10.54%
3	Portugal	5.49%
4	Australia	4.70%
5	Brazil	4.53%

By Affiliation

Rank	Affiliation	Share of listed set
1	Chinese Academy of Sciences	3.85%
2	CNRS (Centre National de la Recherche Scientifique)	2.43%
3	Universidade de Lisboa	2.22%
4	Ministry of Education of the PRC	2.14%
5	Faculdade de Ciências da Universidade de Lisboa	1.98%

By Subject Area

Rank	Subject Area	Share of listed set
1	Environmental Science	33.33%
2	Agricultural & Biological Sciences	23.78%
3	Earth & Planetary Sciences	18.18%
4	Social Sciences	4.42%
5	Engineering	3.50%

This curated dataset now serves as the foundation for the next stage of the project—developing and testing the LLM-assisted pipeline to automatically extract study locations and thematic areas from publication metadata.

B) LLM Pipeline Development and Deployment

The large language model (LLM) pipeline has been successfully tested and deployed. Built on the Qwen2.5 instruction model and supported by a high-performance infrastructure (4× RTX 4090 GPUs), the system automates structured data extraction from the literature. Specifically, it identifies and records study areas (including site, country, and regional information) with associated precision and confidence flags, as well as thematic classifications (primary and secondary). Quality assurance is maintained through deterministic prompting and automatic

flagging of entries requiring review, particularly those with low-confidence scores or multiple study sites.

C) Research Visit and MSc Supervision

A research visit to the University of Glasgow facilitated on-site alignment with project partners and the supervision of an MSc thesis directly linked to the LLM pipeline and its Scottish coastal application. During this period, standard operating procedures (SOPs) were established, and a pilot list of Scottish coastal segments was agreed upon to guide subsequent analyses.

Fieldwork and engagement activities were conducted in collaboration with ongoing coastal erosion projects. The Dundee site was surveyed. Several candidate shoreline sections were shortlisted for cross-referencing literature-based evidence with observed erosion data.

Some project outcomes were shared at the CUPUM conference, where a dedicated workshop on LLM-assisted evidence mapping attracted interest from the wider urban and environmental analytics community. These engagements helped seed new collaborations and opened opportunities for data-sharing across institutions.

D) Initial Results

The operational database schema has been finalized, integrating core fields such as DOI, title, abstract, authors and affiliations, normalized site and theme identifiers, and associated confidence flags. Early pilot testing of the LLM shows robust performance in detecting study countries, with regional and local identifications flagged for human verification where ambiguity remains. The Glasgow-based supervision model is now established, and emerging partnerships—particularly with coastal erosion researchers—are providing access to shoreline datasets and validation pathways for the next stage of analysis.

Next Steps (3-6 Months)

Over the next three to six months, the project will identify future coastal human-activity hotspots and assess equity gaps between scholarly attention and projected population exposure. Using open, spatially explicit datasets—including UN World Population Prospects, WorldPop, and GHS-POP—we will map population projections for 2030, 2040, and 2050 within defined coastal buffers and construct a Human Activity Index (HAI). LLM-derived study sites will be converted into attention-density surfaces, enabling equity analysis through coverage ratios, Gini-based inequality metrics, and an "under-served" index highlighting high-population but under-studied zones. Outputs will include global and Scottish-scale bivariate maps, Lorenz curves, and dashboard prototypes, supporting policy discussions on where future monitoring, research, and synthesis efforts should be prioritized.

Visiting Fellowship costs - £4,979.49