Internship (M1/M2): Large language models for ecological information extraction

Duration: 6 months

Location: Laboratoire d'Écologie Alpine, Grenoble, France

Supervision: Nicolas Le Guillarme (research engineer), Wilfried Thuiller (senior researcher)

Contact: nicolas.leguillarme@univ-grenoble-alpes.fr

Aim of the project

Understanding community structure and dynamics is a key element of modern ecology, especially in the light of global change. Whether there are species-based or trait-based, approaches that aim at improving our understanding of the assembly of communities and their impact on the functioning of ecosystems require extensive information on the organisms that make them up. This includes information about organism traits and roles, which represent their physiological, morphological, or life-history characteristics, and the interactions they have with other organisms and the environment. There are several open-access databases that centralize some of the available knowledge on organism traits and interactions. However, most of the information remains dispersed in unstructured form throughout scientific and grey literature, making it challenging to use as part of large-scale and multi-taxa biodiversity studies. **The goal of this project is to develop NLP tools to automatically extract information on organism traits and interactions from textual documents to complement existing databases.**

Large Language Models (LLMs), such as GPT, have demonstrated a revolutionary ability to retrieve and analyze natural language, including its context and distinct nuances in meaning, and transform the data, potentially delivering it in structured forms. These capabilities hold the promise of LLMs as a comprehensive tool for extracting and structuring information from textual data in ecology (Castro et al., 2024). However, it is also apparent that their performance in these tasks depends on the models used, the type of data, and the desired structure of the extracted information. The aim of this internship is to evaluate the capacity of these models for information extraction in ecological research. More specifically, we aim to determine whether it is possible to build robust ecological information extraction models by directly prompting LLMs.

Activities

The work will consist in the following task:

1. Create a gold-standard corpus for organism trait and/or interaction extraction.
2. Design and implement a few-shot learning prompt-based method for taxonomic entity recognition and organism trait/interaction extraction.
3. Compare the performance of the proposed method with that of TaxoNERD (Le Guillarme and Thuiller, 2022) for taxonomic entity recognition.
4. Compare performance of different LLMs (e.g. GPT, LLAMA 2, MISTRAL 7B) on the two downstream tasks.

Profile sought

A student in the last years of an engineer’s or research master's degree (M1-M2) or in a gap year, specializing in applied mathematics or artificial intelligence. The candidate must have a sound knowledge of machine learning (deep learning, reinforcement learning) and good Python programming skills. Previous experience in text mining, information extraction, NLP or LLMs would be appreciated. The candidate should also be able to demonstrate autonomy and good communication skills.

References

Castro, A., Pinto, J., Reino, L., Pipek, P., & Capinha, C. (2024). Large language models overcome the challenges of unstructured text data in ecology. *bioRxiv*, 2024-01.

Le Guillarme, N., & Thuiller, W. (2022). TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, *13*(3), 625-641.