POLITÉCNICA

E.T.S. DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS

Título de las prácticas (Title of the internship):

Exploring Gene-Disease Associations through the use of Large Language Models

Descripción de las funciones del alumno (Description of the student´s tasks)

The research project is dedicated to exploring the potential of Large Language Models (LLMs) to uncover hidden relationships between genes and specific groups of diseases. This undertaking aims to contribute valuable insights to the fields of genetics and biomedicine. Within the vast expanse of biomedical literature, a wealth of knowledge remains undiscovered and ready for exploration. However, the challenge lies in extracting essential insights from this extensive body of publications. To address this challenge effectively, a subset of diseases of particular interest will be carefully chosen, based on their significance in current biomedical research. The work aims to apply LLMs for such task. In this context, we aim to explore the application of LLMs from three perspectives: 1) the application of generalist LLMs such as Google Bard or ChatGPT; 2) the application of specific ones such as DoctorGPT or any other biomedical-based LLM; and 3) the fine-tuning or adaptation of other open LLMs such as LLAMA2, to be fine-tuned using specific biomedical content.

Requisitos (Prerequisites): *(indicar titulación y curso) (give Grade and academic year); otros requisitos adicionales (idiomas, informática, otros conocimientos, etc) (other aditional prerequitistes (languages, informatics, other knowledge, etc)*

- Candidates should hold a bachelor’s degree in biotechnology, biomedical engineering, computer sciences, or related field.
- Strong programming skills are considered a valuable asset.
- Proficiency in English is a mandatory requirement, as the analysis of scientific literature and communication predominantly occurs in English.
- Previous knowledge about NLP or LLMs is encouraged.

Proyecto formativo (Training Project)

Module EXTERNAL PRACTICES. The fundamental goal of the external practices is to guide the student in applying his previously acquired knowledge to real tasks in a group work environment that realistically represents the work conditions the students will encounter in their future roles. In this way, the student will be able to get familiar with a working environment (work schedule, responsibility, attitude, organization, etc), and with adequate working methodology in professional reality, contrasting and applying the acquired academic knowledge.

This training project gives students a special chance to combine their interests in genetics, biomedicine, and NLP/LLMs. It helps them to grow as researchers in these fields by teaching them the skills to discover valuable insights from the scientific literature.

-Supervision and Mentoring: the mentor or advisor will guide and support the student throughout the project. It will offer regular feedback, answer questions, and facilitate learning.

-Duration and Schedule: It includes a starting date and end date, and establishes a flexible schedule that accommodates the student's availability.

-Learning objectives: Development of LLM prompting engineering, understanding of gene-disease relationships, and ability to contribute knowledge generation in the field.

Actividades a desarrollar en la práctica académica (Activities that will be performed in the academic internship):

1. Data collection: Collect a comprehensive corpus of scientific literature related to genetics, genomics, and the specific set of diseases of interest from reputable sources and databases. Ensure that the data is in a suitable format for NLP analysis.
2. Data preprocessing: Preprocess and clean the textual data to make it suitable for LLM analysis. This includes the preparation of potential prompts.
3. LLM access development: Develop strategies based on the creation of code to automatically query different type of LLMs to process the result.
4. LLM deployment and tuning: Perform work to deploy local LLMs such as LLAMA2 and prepare a corpus of data for its optimization in the domain of the work.
5. Interpretation and visualization: Interpret the results of the LLM outputs and visualize the insights using appropriate tools and techniques.
6. Evaluation and validation: Asses the accuracy and precision of the relationship extraction methods using benchmark datasets.
7. Knowledge Generation: Throughout the entire process, it will continually assess and validate the knowledge generated from the LLM application process. This ongoing evaluation ensures that the insights and information extracted contribute substantively to the project's objectives, promoting an adaptable approach to refining and discovering new knowledge.
8. Documentation and Reporting: Document the methodologies, findings, and insights generated from the analysis. Prepare a report summarizing the project's outcomes and contributions.

Nº de plazas: **(Nr. of places)**	1

¿El alumno tendrá trato habitual con menores? **(Has the student dealings with underage persons?)**	No
Fecha de inicio: **(Starting date)**	08/01/2024
Fecha de fin: **(End date)**	31/05/2024
Horas semanales: **(Weekly hours)**	
Horario jornada laboral: **(Working hours)**	
Importe Ayuda/Bolsa de estudio: **(Amount of fellowship / remuneration)**	**- €/mes**
Tutor académico: (Academic tutor (UPM)) Email:	Alejandro Rodríguez González alejandro.rg@upm.es
Departamento tutor académico: **(Dept. of academic tutor)**	Lenguajes y Sistemas Informáticos e Ingeniería de Software

Tutor empresa: **(External tutor)**	Paloma Tejera Nevado
Email tutor empresa: **(Email external tutor)**	paloma.tejera@upm.es
Departamento tutor empresa: **(Dept. of external tutor)**	Medical Data Analytics Laboratory (MEDAL)
Ubicación de la estancia de las practicas **(Location of the internship)**	Parque Científico y Tecnológico de la UPM, Crta. M40, Km. 38, 28223 Pozuelo de Alarcón, Madrid
ENTIDAD COLABORADORA: **(Collaborating Entity)**	CTB (CTB-UPM) Center for Biomedical Technology
A cumplimentar por Oficina Prácticas ETSIAAB: **Créditos a reconocer (Nº ECTS)**:	

Enviar por email a: OFICINA DE PRÁCTICAS ACADEMICAS EXTERNAS – ETSIAAB
secretaria.pei.etsiaab@upm.es – Secretarias: Visitación Pérez / Susana Pardo - Tfno: 913363686)