

Plan: Erasmus Mundus Master in Language and Communication Technologies (LTC)

Subject: Practicum (Internship i.e. prácticas obligatorias)

DETAILS OF THE COMPANY

Nombre de la empresa:	HiTZ Basque Center for Language Technology
Persona de contacto	
Email de contacto	emlct.internship@ehu.es
Teléfono de contacto	
País	
Provincia	Gipuzkoa
Localidad	Donostia

CONTACT DETAILS OF THE TUTOR: the supervisor within the university

Given name	Aitor
Family name	Soroa
Email	a.soroa@ehu.eus

DETAILS OF THE INTERNSHIP

Title	Efficient methods to learn generative language models to follow instructions in multiple languages
Goal	The goal of this project is to use parameter efficient methods to learn generative large language models (LLM) to follow human instructions in various languages. The student will dive into the so called instruction learning phase of LLMs, and apply parameter efficient techniques that do not require to update all the parameters of the model. The work will be carried out within the DeepR3 project.
Tasks	The aforementioned goals require fulfilling the following tasks: <ul style="list-style-type: none">● Task 1. Study and select generative language models that are already pre-trained in the languages of interest.● Task 2. Select available instruction datasets.● Task 3: Translate the instruction datasets to the target languages.● Task 4: Use parameter efficient methods such

	<p>as adapters or LoRA to teach the base language models to follow the instructions.</p> <ul style="list-style-type: none"> ● Task 5. Evaluate the models using a suitable testbed.
Learning outcomes	<p>By the end of this internship, the students will</p> <ul style="list-style-type: none"> ● Learn to use and adapt generative language models to new tasks. ● Use machine translation systems to translate the instruction datasets. ● Learn parameter-efficient methods to adapt language models to the new tasks. ● Learn to evaluate generative language models using prompts.
Materials /Resources	The student will use resources from the HiTZ centre, including computing power (GPU), etc.
Starting date:	
End date:	June (due date for transcript of records in GAUR)
Timetable:	Flexible timetable and work schedule.
Number of hours (10ECTS):	250h
Language	The internship will be developed in English. Speaking other languages (specially Spanish) is recommended but not necessary.
Financial support	0€
Intellectual Property %	<ul style="list-style-type: none"> ● The work done through the project will be publicly accessible.
Specific requirements (background of the candidate)	<p>Background of the candidate:</p> <ul style="list-style-type: none"> ● Preferably EMLCT Y2 student ● Preferable Engineer or Computer Scientist. ● Background on Deep Learning and LLMs. ● Good programming skills in Python.