Short Description of Internship and Thesis Activities, and Expected Outcome:

**TITLE:**
Building Italian chat-bot for IT Help Desk by using transfer learning from ChatGPT and similar

**DESCRIPTION:**
The project aims to build a custom-made chat-bot. Such a chat-bot should leverage the Information Technology (IT) Help Desk's data-set owned by GPI. The IT Help Desk is specialized in troubleshooting computer-related problems experienced by GPI's clients. Since the Help Desk is already running from several years by now, it has cumulated more or less 750000 records (called "tickets") about either common or less common computer-related problems. All tickets are written in Italian. Each ticket may be an email exchange between client and IT expert or a very short summary of the problem written by the IT technician.

**METHOD:**
The idea was to use an open source Large Language Models (LLMs) like LLaMA (Touvron et al. 2023) or GPT4All etc. as a backbone. By fine tuning/transfer learning the LLM will then be fed with data coming from the Italian data-set.

**HURDLES TO OVERCOME:**
- test whether the LLM generalizes to Italian data;
- although some tickets have a written step-by-step solution to the problem, some tickets are very succinct (thus, no description of the steps undertaken for solving the problem);
- although the interaction between client lamenting a problem and IT technician may happen as a telephone call, no calls were recorded. This means that a step-by-step solution taught by phone by the IT technician to the client is lost. Therefore we have to leverage the textual data only;

**GOAL:**
The ideal chat-bot should solve by itself at least 10 % of calls to the IT Help Desk. Indeed, we believe that at least 10 % of calls are simple and redundant IT problems (e.g. how to install new driver for the printer). If this works, this can save resources to GPI since some IT technicians of the Help Desk may work for other departments, instead of being busy at the Help Desk.

SCHEDULE (ROUGH ESTIMATE):
- month 1: read literature about anonymization and pseudo anonymization (see link below for an example); find and test optimal and GDPR-compliant software or procedure for anonymizing data-set;
- month 2: pre-processing data-set; find the right open source LLM;
- month 3: build the Deep Learning model (by transfer learning/fine-tuning) and train it;
- month 4: further iterations of the model and further training; test and validate the whole pipeline with unseen data;
- month 5: once the pipeline is stable, provide some help to developers with the deployment;
- month 6: writing thesis, documentation and polish code;

REFERENCES:

**Required Candidate Skills and Prerequisites:**

**SKILLS:**
- experience with Python programming (e.g. TensorFlow, PyTorch), specifically with an emphasis on Transformers and transfer learning;
- some knowledge of the scientific literature about Large Language Models (e.g. ChatGPT, GPT4All, LLaMA etc.);
- some understanding of Linux systems (for getting computational power) and version control (e.g. GitHub, GitLab etc.);
- solid understanding of statistics;
- good English skills.