Title of the internship: A Multi-task End-to-end Language model for Intelligent Argumentation (AMELIA)

Target student & Duration: M2 student (DISS or AI) - 6 months

Description of the internship: We aim to provide an open-source and lightweight LLM which can map texts to semi-structured argumentation graphs efficiently across multiple domains and languages. The idea is to leverage a small set of models as the backbone for the multiple required tasks, including: (i) argument recognition, (ii) relation and fallacy inference, (iii) argument quality assessment, and (iv) argument re-writing.

We will first collate the data for the different tasks (using open-source research data, online or synthetic annotation tasks, and anonymized debates resources), forming the largest multi-task argumentation dataset. This dataset will be crucial for this project and serve as the basis for the evaluation of the argumentation capabilities of future LLMs. To mitigate benchmark contamination, we will put in place methods to dynamically update the dataset, inspiring ourselves from MixEval[1]. Second, we will select a set of suitable open-source LLMs, focusing on the best performing ones with less than 15 billion parameters (including Meta's LlaMA 3.2, Microsoft Phi-3, Google's Gemma, Mistral NeMo, etc.). For the training of the model(s), we plan to re-use the approach of BioMistral[2], combining a pre-training phase on the Jean Zay HPC using argumentation research documents in the Commercial Use Allowed subset with a supervised fine-tuning phase on our multi-task argumentation dataset. Multiple quantized versions of the trained models will be uploaded on HuggingFace to democratize them and enable their execution on smaller devices. As the proposed language model is intended to work within a DPPD, it will require new methods to perform retrieval-augmented generation (RAG) on argumentation graphs. Contrary to knowledge graphs where the relation between entities is highly expressive, argumentation graphs often only have a few number of relations (e.g., attacks or supports), making the retrieval for complex queries more challenging. We plan to create an automatic converter which will construct an augmented knowledge graph from an argumentation graph, allowing for efficient RAG on argumentation graphs.

Implication of the supervisor: Bruno Yun (teaching for 155.5 hours in 2024-2025). The teaching load is split as follows:

- INF3007L (LIFAPOO L3): 27h
- INF3051L (Projet Informatique L3): 16h
- INF2347M (Dynamique des connaissances M2 IA): 37.5h **(course lead)**
- INF2511M (Theory and Applications of Large language models M2 IA): 39h **(course lead)**
- INF2494M (Machine learning techniques and applications M2 DISS): 15h
- INF1103M (Technique de l'intelligence artificielle M1): 12h
- INF1217M (Tuteur stage informatique M2): 9h

Significance for the Computer Science Department: AMELIA represents a strategic opportunity for the Computer Science Department by strengthening research in LLMs, argumentation, and automated reasoning. It promotes the development of open-source tools accessible to both the academic and industrial communities while involving Master's students in innovative challenges aligned with their coursework. Additionally, the project highlights the department's expertise in LLMs, and aligns with CNRS priorities, thereby enhancing the department's scientific visibility and impact.

Historique des attributions de gratifications:

•Open position for HAGARICE project, M2 student, 6 months - CNRS Transversal

•Elliot Faugier, M2 IA, 6 months in 2024 - the University Lyon 1, AAP Accueil.

---

[1] Ni et al. "MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures". In: CoRR abs/2406.06565 (2024). arXiv: 2406.06565

[2] Labrak et al. "Biomistral: A collection of open-source pretrained large language models for medical domains". In: arXiv preprint arXiv:2402.10373 (2024).