# Junior Cedric Tonga

+971 55602 3526 | junior.tonga@mbzuai.ac.ae | linkedin | github | **My website** | Abu dhabi, UAE

***Interests***: *NLP in general, especially Low-resource NLP, Multilingual NLP, cultural adaptation, applications of NLP (education, finance).*

## Education

**PhD student in NLP, MBZUAI** Abu dhabi, UAE
*Research on multilingual, low-resource NLP, reasoning* *Aug. 2025 – Aug. 2029 (expected)*

**Master 2 MVA (Mathematics, Vision, Learning), ENS Paris-Saclay; CGPA: 4/4** Gif-sur-Yvette, France
*Computer vision, Speech recognition, Advanced learning for text and graph data, Time Series, ML, DL* *Oct. 2023 – Apr. 2024*

**Master 1 in artificial intelligence, Université Paris-Saclay; CGPA: 3.7/4** Gif-sur-Yvette, France
*Applied statistics, maths for data science, NLP, optimization, big data (Hadoop...), ML, DL* *Sept. 2022 – April 2023*

**Bachelor's degree in computer science, Faculty of sciences of Gabes; CGPA: 4/4** Gabes, Tunisia
*Data structure and algorithmic, Probability and Statistic, Linear algebra, programming* *sept. 2019 – June 2022*

## Experiences

**MBZUAI, NLP department** Abu dhabi, UAE
*Research associate II* *October 2024 – in progress*

- *Supervisor: Dr Fajri Koto*
- Currently working on two projects with my supervisor, aimed for submission to **EMNLP 2025**
- Current Status: Ongoing research and experimentation
    - *Lead, project 1: How far the best LLMs can be used to extract cultural commonsense knowledge graph ?*–our goal is to build a cultural commonsense knowledge graph (CKG) that have geographical context.
    - *Lead, project 2: Multilingual reasoning in Education*

**EvidenceB, AI Research team** Paris, France
*NLP research scientist intern* *April 2024 –August 2024*

- *Supervisors: Dr Pierre-Yves Oudeyer and Dr Benjamin Clement*
- Exploring the use of LLMs (GPT-4o and Llama-3-8B-instruct) as teachers to generate effective hints for students simulated through LLMs (GPT-3.5-turbo, Llama-3-8B-Instruct, Mistral-7B-instruct-v0.3) tackling math exercises designed for human high-school students, and designed using cognitive science principles.
- Identified common student errors, developed prompts to aid self-correction, and compared teacher models performance using best prompt in generating pedagogically effective hints.
- Results revealed that effective hints improved student model performance, especially GPT-3.5-turbo at lower temperatures. Llama-3-8B-Instruct as a teacher showed better overall performance than GPT-4o and Mistral-7B-Instruct showing decreased accuracy at higher temperatures.
- **Paper accepted at NeurIPS 2024 FM-Assess Workshop (poster): Automatic Generation of Question Hints for Mathematics Problems using Large Language Models in Educational Technology.**

**Université Quebec à Montréal, CIRST lab** Montreal, Canada
*NLP research intern* *June 2023 – Aug.2023*

- *Supervisors: Dr Marie-Jean Meurs and Dr Diego Maupomé*
- Conducted a study to assess the robustness of monolingual and multilingual language models to specific linguistic structures within the context of suicide prevention tools aimed at accommodating the cultural diversity of individuals in distress. Trained XLM-R, distiluse-base, and CamemBERT-base models on a dataset of French sentence pairs using the simple contrastive learning of sentence embeddings(SimCSE) method.
- Results revealed that while pre-trained multilingual models initially performed well, post-training, monolingual models demonstrated superior performance over multilingual models.
- **Work presented at the ACFAS congress in Ottawa in may 2024.**

**Digital Research Center of Sfax, Brain4ICT Team** Sfax, Tunisia
*NLP research intern* *Feb. 2022 – May 2022*

- *Supervisor: Dr wael ouarda*
- Researched, collected, and cleaned product review data in Francamglais Cameroonian dialect from YouTube via Python script. Utilized BERT for sentiment analysis, achieving an 86% accuracy post fine-tuning. Deployed the model on a web application using Flask for user sentiment analysis.
- **Paper accepted at IWCMC 2024: AfriDial: African Dialect Model based on Deep Learning for Sentiment Analysis.**

## Publications

Abdelrahman, S., **Tonga, J. C**., Khalid, A., Saeed, A., Farah, A., Chatrine, Q., Karima, K., Sara, S., Yaser, A., Fajri, K. (2025). Commonsense Reasoning in Arab Culture. **Under review at ACL 2025**

**Tonga, J. C**., Clement, B., Oudeyer, P. Y. (2024). Automatic Generation of Question Hints for Mathematics Problems using Large Language Models in Educational Technology. **NeurIPS 2024 Workshop on Large Foundation Models for Educational Assessment (FM-Assess). Published by Proceedings of Machine Learning Research (PLMR).**.

Sassi, A., **Tonga, J.**, Poaty, S., Steve, S., Abakar Adjid, D. I., Cherif, M., Ouarda, W. (2024). AfriDial: African Dialect Model using Deep Learning for Sentiment Analysis. **International Wireless Communications and Mobile Computing (IWCMC) 2024**.

## ACHIEVEMENTS & AWARDS

**G-Research Paris Quant Challenge 2024** Paris, France
*First-place winner team (2 participants: myself and my teammate) of the G-Research Paris Quant Challenge 2024.* *2024*

**SaclAI school excellence scholarship (MixtAI)** ENS Paris-saclay, France
*Paris-Saclay awards €10k to top AI Master's students for academic excellence* *sept. 2023–June 2024*

**Globalink Research Internship fellowship-MITACS(**$\simeq \$10kCAD$**)** UQAM, Canada
*12-week internship program at a Canadian university.* *may 2023– august 2023*

**Idex international internship grants(IDEX)** Université Paris-saclay, France
*internship bursary awarded to international interns on the basis of academic results* *may 2023– august 2023*

**SaclAI school excellence scholarship (MixtAI)** Université Paris-saclay, France
*Paris-Saclay awards €10k to top AI Master's students for academic excellence* *sept. 2022–June 2023*

**Research Excellence Award (resigned)** Université sorbonne paris-nord, France
*€12k in the first & €10k in the second year of EUR Msc.* *sept.2022-sept.2024*

**Hatem Ben Taher Award** FSG, Gabes, Tunisia
*Best student in bachelor's degree in computer science of all 15 schools of the University of Gabes.* *June 2022*

**Tunisian government scholarship** FSG, Gabes, Tunisia
*cooperation scholarship between the Cameroonian and Tunisian states.* *sept; 2019- Aug. 2022*

## Service

Primary reviewer, COLING 2025.

## Academic projects

**Molecule Retrieval with Natural Language Queries** | *Hugging Face, pytorch* Feb.2024

- Participated in a team in the challenge aimed at identifying molecules(represented as graphs) corresponding to given textual query. Our general approach comprises four blocks: text and molecule encoding, modality alignment, and retrieval using SciBERT, GTN, GPS and others models. By integrating various loss functions and exploring training strategies, We acheived a rank of 7 out of 52 teams.

**Model compression using knowledge distillation and quantization** | *Python, Unet* Feb. 2024

- The project's goal was to distill a Unet model for groove segmentation using knowledge distillation and quantization, implementing it manually without relying on PyTorch APIs.

**Lymphocytosis classification** | *Hugging Face, sk-learn, pytorch* March 2024

- Participated in a team in the challenge focusing on binary patient classification (reactive or malignant), we adopted a multimodal strategy. This involved crafting attribute-based and ResNet-based image models with aggregation methods, alongside employing Multiple Instance Learning incorporating a custom aggregation inspired by focal loss. We finished 2nd out of 39 Teams.

**Entity detection and relation extraction** | *Python, prodigy, NER, trankit, spacy, git, transformers* April 2023

- The objective was to extract the entities of a patent specific to a given domain by using BERT and to find the relationships between these entities in order to create a knowledge graph.

## Technical Skills

**Languages**: French, English
**Programming Languages** : Python (advanced), Java (prior experience), C/C++(prior experience), SQL(prior experience).
**Developer Tools & framework**: Git, Docker, VS Code, Spark, hadoop, Linux, Latex
**Libraries**: PyTorch, Numpy, Pandas, Scikit-learn, Matplotlib, Seaborn, NLTK, SentenceTransformers, transformers
**Other** : High-Performance Computing, Hugging Face