**Exploration of the Use of Large Language Models (LLMs) for Occupational Code Classification.**

**Project background**

Large Language Models (LLMs) have shown significant potential in natural language processing, enabling advancements in automated tasks like occupational code classification. Traditional methods of classifying occupations, such as International Standard Classification of Occupations (ISCO), International Standard Classification of Education (ISCED), International Standard of Industrial Classification (ISIC), rely on manual input, which can be inefficient and error-prone. By using LLMs, job titles, descriptions, and Educational can be automatically analyzed and assigned accurate occupational codes, improving efficiency and scalability.

This project explores the use of LLMs for occupational code classification within Rwanda’s labor market. It aims to enhance the accuracy and speed of classification, providing valuable insights for labor market analysis and supporting informed policy-making. As part of NISR’s Big Data and Data Revolution department, which is responsible for fostering innovations, this project will contribute to the ongoing advancements in data-driven solutions.

**Problem Statement**

Traditional methods of occupational code classification are manual, time-consuming, and error-prone, making it difficult to process large volumes of data efficiently. This is particularly challenging in Rwanda’s evolving labor market, where emerging job roles require more adaptive classification systems. This project explores the use of Large Language Models (LLMs) to automate and enhance the accuracy, speed, and scalability of occupational classification, aiming to provide better insights for labor market analysis and policy development.

**Methodology**

The project will use advanced data science techniques such as:

* Text preprocessing (cleaning, tokenization, and embedding generation, all these can be applied on both traditional and LLM based methods).
* LLM selection **(**we are still searching on which LLM model that can be used)
* Model validation
* Use RAG with LLMs to capture context and relationships in textual data.
* Deployment

**Expected Outputs:**

1. A functional **AI-powered code classification tool** that will generate codes based on descriptions.
2. Documentation of the model development and implementation process.

**Impact:**  
This project will:

* Streamline NISR’s classification workflows.
* Promote scalability and transferability of the tool to other departments requiring automated classification.
* This will help in time saving and rapid data collection.
* It will also help in minimizing codification error

**Project Timeline**

| **Week(s)** | **Phase** | **Key Activities** |
| --- | --- | --- |
| 1–2 | Project Planning and Setup | - Define objectives, scope, and deliverables.  - Identify and gather relevant datasets.  - Set up tools and environment. |
| 3–5 | Data Collection and Preprocessing | - Collect datasets from sources.  - Clean and preprocess data.  - Perform exploratory data analysis (EDA). |
| 6–8 | Feature Engineering and Model Selection | - Prepare text features.  - Select LLM and baseline models for comparison. |
| 9–11 | Model Training and Evaluation | - Train LLM on occupational data.  - Evaluate performance using metrics. |
| 12–13 | Optimization and Final Model Selection | - Fine-tune hyperparameters to improve performance. |
| 14–15 | Deployment and System Development | - Develop a pipeline for automated classification.  - Test and validate the system with new data.  - Document and prepare for deployment. |
| 16 | Submission and Wrap-Up | - Present findings and demonstrate the tool.  - Reflect on outcomes and future improvements. |