## **Meeting Notes**

### **Date & Participants:**

* **Date**: 28th March 2025
* **Participants**: Hao Chen, Masoud Rahimi, Ming Chen, Pranav Pai, Chaoge Zhu, Haoran Guo, Yang Jin, Yushi Wang.

### **Introduction:**

* Masoud Rahimi (Lead Data Scientist, AURIN) and Hao Chen (Senior Data Scientist, AURIN) provided an overview of the company and project requirements.
* AURIN supports researchers by providing data, analytics, and infrastructure tools.

### **Project Overview:**

* **Goal**: Develop a data validation proof-of-concept to improve AURIN's existing ETL pipeline.
* **Main Challenges**:  
  + Detecting anomalies, missing data, biases, and data inconsistencies.
  + Current methods are rule-based and insufficient due to diverse datasets.

### **Expectations & Requirements:**

* **Scope**:  
  + Create a standalone Proof-of-Concept (POC), not integrated into AURIN's existing ETL pipeline at this stage.
  + Explore data validation methods using AI, particularly large language models (LLMs).
* **Technical & Data Aspects**:  
  + Use static datasets (no streaming).
  + Small-scale prototype suitable for local execution (minimal GPU usage).
  + Allowed to use external LLM APIs (e.g., OpenAI) or locally-hosted LLMs.
  + Clearly justify choice of technical tools and methods.

### **Timeline & Deliverables:**

**Semester 1 (Planning & Initial Analysis):**

* **Mid-April**: Initial project timeline & scope definition.
* **End of May**: Data collection and comprehensive literature review.
* **Mid-June**: Data exploration, refined project plan, and initial findings presentation (in-person at AURIN).

**Semester 2 (Implementation & Reporting):**

* **Late July - Mid-September**: Complete implementation.
* **Late October**: Final oral presentation at AURIN, submission of reports, and all deliverables (scripts, data, prototypes).

### **Important Guidelines:**

* Maintain confidentiality of AURIN’s data.
* Use clean coding practices; provide well-documented, maintainable code.
* Manage code collaboratively via GitHub.
* Keep data securely (recommended: OneDrive, University infrastructure).
* Regular communication: fortnightly meetings, Slack channel for internal discussions.

## **Next Steps & KPIs**

### **Immediate Actions (Next 2 Weeks):**

1. **Set Up Regular Meetings**:  
   * Agree on three potential fortnightly meeting times.
2. **Initial Project Timeline**:  
   * Define clear initial project milestones and tasks by mid-April.
3. **Data Acquisition Plan**:  
   * Identify potential sources (AURIN data, external datasets).
   * Clarify specific data formats and requirements.

### **Short-term KPIs (Next 1-2 Months):**

* Complete detailed **literature review** on state-of-the-art data validation methods by **end of May**.
* Acquire and secure necessary datasets for validation tasks by **end of May**.
* Conduct preliminary **Exploratory Data Analysis (EDA)** by **mid-June**.

### **Semester-long KPIs (Next 3-6 Months):**

* Implement and deliver a functional **proof-of-concept** for data validation.
* Prepare comprehensive **documentation**, including:  
  + Justifications for technical decisions.
  + Identified project limitations.
  + Suggestions for future enhancements.
* Deliver an effective **final presentation** demonstrating project results and learnings in **late October**.

These meeting notes and KPIs clearly structure the project's current understanding, future actions, and measurable outcomes.