**Capstone Project Concept Note and Implementation Plan**

**Project Title: AI-Powered Learning Recommender for Refugee Education**

**Team Members**

1. Mesfin Mulugeta Wetere ([msfnmulgeta@gmail.com](mailto:msfnmulgeta@gmail.com))
2. Makda Workeshet Bizuneh ([makdabizuneh2005@gmail.com](mailto:makdabizuneh2005@gmail.com))
3. Ekram Kedir Sherif ([ekedir37@gmail.com](mailto:ekedir37@gmail.com))
4. Meron Birhanu ([meron9803@gmail.com](mailto:meron9803@gmail.com))
5. Anasimos Tesfaye Gebeyehu ([anasimostesfaye@gmail.com](mailto:anasimostesfaye@gmail.com))

**Concept Note**

**1. Project Overview**

Education for refugees is a pressing challenge, with many facing language barriers, limited resources, misaligned resources and a lack of personalized learning options. Our project aims to build an AI-powered recommender system that suggests free online courses (e.g., from Coursera, Khan Academy) tailored to refugees’ profiles, including their language, education level, and career goals. By leveraging machine learning, we seek to make education more accessible and relevant, directly supporting **SDG 4 (Quality Education)**, while also contributing to **SDG 10 (Reduced Inequalities)** and **SDG 8 (Decent Work & Economic Growth)** by enhancing employability through upskilling.

Design for offline use, our system ensures usability in low-resource settings like refugee camps, and address biases to promote fairness. By personalizing learning, we hope to boost course completion rates and help refugees gain skills for better futures.

**2. Objectives**

* Build a hybrid AI model combining collaborative filtering and NLP to recommend personalized courses.
* Increase access to free, high quality education and promote long-term socio-economic empowerment.
* Enable offline access to recommendations for users in low-bandwidth environments.
* Minimize biases (e.g., gender bias in STEM courses) to ensure equitable recommendations.
* Align course suggestions with vocational skills to enhance employability.
* Provide multilingual and skill-aligned learning resources tailored to each refugee’s background and goals.
* Ensure ethical and inclusive design that respects privacy and supports low-resource settings.
* Align with global development goals, especially those targeting education, equity, and work opportunities.
* Achieve a 30% improvement in course completion rates among refugee learners.

**3. Background**

Refugees often lack access to education due to disrupted schooling, language barriers, inadequate infrastructure and also face difficulties adapting to new learning environments. Traditional systems are ill-equipped to address their unique linguistic, cultural, and technological needs. However, the growth of online educational platforms and advancements in artificial intelligence present an opportunity to create customized educational pathways.

Previous efforts, like Coursera for Refugees, provide free MOOCs but suffer from high dropout rates (40%, UNHCR 2021) because they aren’t tailored to individual needs. Other initiatives, like Kolibri, offer offline content but lack personalization. Even though existing solutions provide basic support they lack advanced natural language understanding.

A machine learning approach is ideal here because it can analyze user profiles and course metadata to deliver precise recommendations, while NLP can interpret complex needs (e.g., “I want to learn nursing in Arabic”). Inspired by recent developments, our project builds on these efforts by combining collaborative filtering for user-based recommendations with Sentence-BERT for semantic matching, ensuring both personalization and offline accessibility. We aim to integrate structured and unstructured data to offer smarter, more context-aware recommendations.

**4. Methodology**

Our system uses a **hybrid machine learning model**:

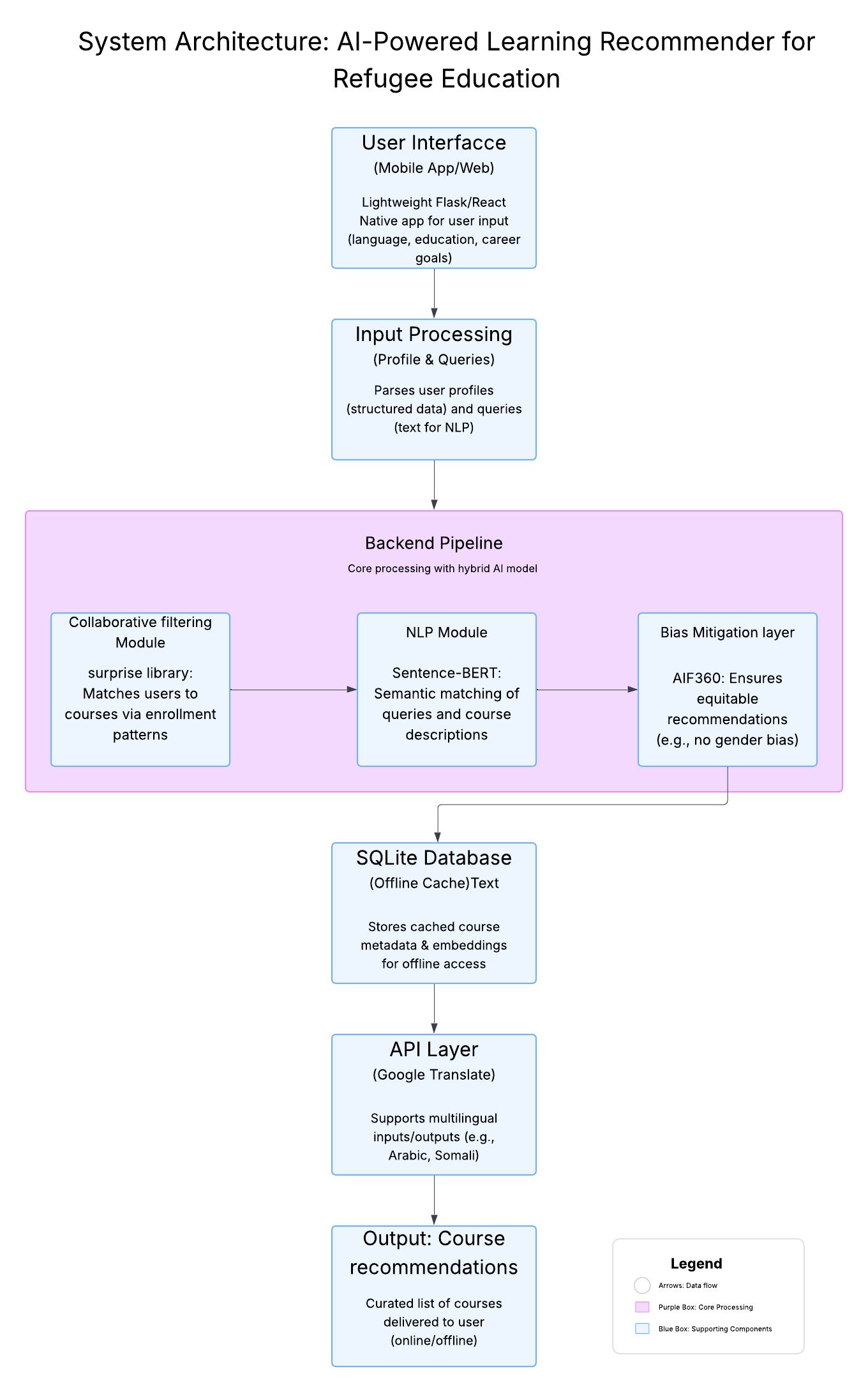
* **Collaborative Filtering**: Using the surprise library, we’ll match refugees with similar learners and courses based on similar users’ completion patterns. This handles structured data like enrollment history.
* **Deep learning (NLP)**: Sentence-BERT will semantically align user inputs (e.g., “healthcare jobs”) with course descriptions, fine-tuned on refugee-generated text for better accuracy.
* **Bias Mitigation**: AIF360 will detect and reduce biases (e.g., over-recommending STEM to males) by debiasing embeddings.
* **Offline Functionality**: Cached course embeddings and an SQLite database will enable recommendations without internet access.
* **Data collection and preprocessing**

**Refugee Profiles:** UNHCR Microdata Library (CSV) – handle missing data, anonymize PII

**Course Metadata:** Kaggle (Coursera/edX) (CSV/JSON) – extract topics, languages, durations

**User Feedback:** Synthetic survey data (Text) – apply sentiment analysis

* **Evaluation and testing**
* Simulate user scenarios using synthetic profiles
* Assess recommendation accuracy via relevance scores
* Collect user feedback for iterative improvements
* The model will be quantized to run on low-end devices, and a TF-IDF fallback will ensure functionality if Sentence-BERT fails in resource-constrained settings. This approach ensures personalization, fairness, and accessibility, building on prior work while addressing gaps in offline deployment and bias correction.
  1. **Architecture Design Diagram**

****

**Diagram Description**:

* **User Interface (Web/Mobile App)**: A lightweight interface (built with Flask or React Native) where users input their profile (language, education level, career goals) and queries (e.g., “healthcare jobs”). It’s optimized for low-end devices.
* **Input Processing**: Parses user inputs, converting profiles into structured data and queries into text for NLP analysis.
* **Collaborative Filtering Module**: Uses the surprise library to analyze enrollment histories and recommend courses based on patterns from similar users. Outputs a ranked list of courses.
* **NLP Module**: Employs Sentence-BERT to compute semantic similarities between user queries and course descriptions, fine-tuned for multilingual refugee contexts (e.g., Arabic, Somali).
* **Bias Mitigation Layer**: Applies AIF360 to detect and correct biases (e.g., gender skew in STEM recommendations), ensuring equitable outputs.
* **SQLite Database**: Stores cached course metadata (skills, duration, language) and precomputed embeddings, enabling offline recommendations.
* **API Layer**: Integrates Google Translate API to support multilingual inputs and outputs, enhancing accessibility.
* **Output**: Delivers a curated list of course recommendations to the user via the interface, available online or offline.

The components interact through a Python-based backend, delivering recommendations via the UI, optimized for low-resource environments.

**6. Data Sources**

We’ll use UNHCR Microdata Library (CSV, 10k rows) for anonymized refugee profiles (language, education level), Kaggle Coursera/edX datasets (CSV/JSON, 500+ courses) for course metadata (skills, duration, language), and synthetic mock surveys (text, 200 responses) to simulate user feedback. Preprocessing includes cleaning missing values, anonymizing PII, and extracting relevant course features (e.g., vocational skills). These datasets are chosen for their relevance to refugee demographics and vocational needs, ensuring our model reflects real-world requirements.

**7. Literature Review**

Prior work highlights the need for personalized education for refugees. UNHCR (2021) used collaborative filtering to improve MOOC enrollment but lacked NLP for nuanced queries, reporting 40% dropout rates. IEEE (2022) applied BERT for course matching, achieving 85% semantic accuracy, but its GPU demands limit deployment in low-resource settings. Zawacki-Richter et al. (2019) exposed gender biases in STEM recommendations, underscoring the need for ethical AI. Our project extends these findings by combining collaborative filtering with lightweight Sentence-BERT, integrating AIF360 for bias mitigation, and prioritizing offline access, as inspired by Kolibri’s offline learning model (Learning Equality, 2023).

**Implementation Plan**

**1. Technology Stack**

* **Programming Language**: Python 3.9
* **Libraries**:
  + *surprise****:*** Collaborative filtering.
  + *transformers:* Sentence-BERT for NLP.
  + *AIF360:* Bias detection and mitigation.
  + *pandas, NumPy:* Data preprocessing.
  + *scikit-learn:* Model evaluation.
* **Frameworks**:
  + *Flask:* Web app backend.
  + *React Native:* Mobile app frontend.
* **Database**: SQLite for offline storage.
* **APIs**: Google Translate for multilingual support.
* **Tools**: VS Code, Jupyter Notebook, Google Colab, Git for version control.
* **Hardware**: Low-end mobile devices for testing, laptops for development.

**2. Timeline**

The project spans 20 weeks, broken into phases with specific tasks. Below is a Gantt chart and task distribution matrix.

**Gantt Chart**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TASKS** | **WEEKS** | | | | | |
| **1-2** | **3-4** | **5-8** | **9-12** | **13-16** | **17-20** |
| **Data Collection & Preprocessing** |  |  |  |  |  |  |
| **System Design & Architecture** |  |  |  |  |  |  |
| **Model Development (Collab.Filt.)** |  |  |  |  |  |  |
| **Model Development (NLP)** |  |  |  |  |  |  |
| **Bias Mitigation Implementation** |  |  |  |  |  |  |
| **Offline Function setup** |  |  |  |  |  |  |
| **Testing & Refinement** |  |  |  |  |  |  |
| **Deployment & Evaluation** |  |  |  |  |  |  |

**Task Distribution Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task** | **Mesfin** | **Makda** | **Ekram** | **Meron** | **Anasimos** |
| Data Preprocessing |  |  |  |  |  |
| System Architecture Design |  |  |  |  |  |
| Collaborative Filtering |  |  |  |  |  |
| NLP (Sentence-BERT) |  |  |  |  |  |
| Bias Mitigation (AIF360) |  |  |  |  |  |
| Offline Setup (SQLite) |  |  |  |  |  |
| Testing & Bug fixing |  |  |  |  |  |
| Deployment & Documentation |  |  |  |  |  |

**3. Milestones**

* **Week 4**: Complete data preprocessing and system architecture design.
* **Week 8**: Functional collaborative filtering model (80% accuracy on test data).
* **Week 12**: Integrated hybrid model with Sentence-BERT and bias mitigation (85% MRR).
* **Week 16**: Beta version with offline functionality tested on low-end devices.
* **Week 20**: Deployed system with final report and impact evaluation.

**4. Challenges and Mitigations**

* **Challenge**: Limited real-world refugee data.
  + **Mitigation**: Use synthetic surveys and partner with UNHCR for additional anonymized data.
* **Challenge**: Sentence-BERT’s high computational needs.
  + **Mitigation**: Quantize model and implement TF-IDF fallback for low-end devices.
* **Challenge**: Gender or language biases in recommendations.
  + **Mitigation**: Apply AIF360 debiasing and fine-tune Sentence-BERT on diverse refugee texts.
* **Challenge**: Connectivity issues in refugee camps.
  + **Mitigation**: Prioritize SQLite and cached embeddings for offline use.

**5. Ethical Considerations**

Our project handles sensitive refugee data, so we’ll anonymize PII per UNHCR guidelines to protect privacy. Bias mitigation is critical, as prior systems over-recommended STEM courses to males (Zawacki-Richter, 2019). We’ll use AIF360 to ensure fair recommendations across genders and languages. The system’s impact on refugees must be positive, so we’ll collect feedback to avoid unintended consequences, like recommending inaccessible courses. Partnerships with NGOs like Learning Equality will ensure culturally sensitive deployment.

**6. References**

* UNHCR (2021). *MOOC Recommendations for Refugees*. Geneva.
* IEEE (2022). *BERT for Educational Resource Matching*.
* Zawacki-Richter et al. (2019). *Review of AI in Education*. Computers & Education.
* Learning Equality (2023). *Kolibri: Offline Educational Resources*.
* Reimers & Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.