🚀 **Project 2. AI Engineer Candidate Project Proposal**
Company: ScaleX Innovation
Project Title: CV Extractor with Open-Source LLMs using Ollama
Duration: 5 Days Mistral OCR
Evaluation Tool: Custom Accuracy Metrics (Field-level Extraction Accuracy)
Candidate Level: Intermediate to Advanced AI Engineer

🧠 **Project Summary**
This project aims to build an intelligent CV extraction system using open-source large language models (LLMs) running on **Ollama**. The system should extract structured information such as name, contact info, education, experience, and skills from CVs uploaded in either **PDF (text or scanned)** formats.

The pipeline must:

- Handle both text-based and image-based PDFs (via OCR)
- Use three different **open-source LLMs** for extraction (e.g., **LLaMA 3**, **Mistral**, **Phi-2**)
- Compare the extraction accuracy of the models against a labeled ground truth
- Provide a simple **web app interface** for file upload and result visualization

🎯 **Project Objectives**

- Integrate **Ollama** to run 3 open-source LLMs locally
- Build a data processing pipeline:
  - For text-based PDFs: use `PyMuPDF (fitz)` or `PyPDF2`
  - For image-based PDFs: use OCR with Visual LLM (e.g. Mistral OCR or others)
- Prompt each LLM to extract structured fields from raw CV content
- Build a basic **Flask** web interface to upload CVs and display extracted results
- Evaluate model extraction accuracy using field-level precision and recall against a labeled sample set of CVs

---

🔧 Technical Requirements

- **LLMs**: Open-source models via **Ollama** (e.g., `llama3`, `mistral`, `phi`)
- **OCR**: with Visual LLM for OCR apps
- **PDF Parsers**: `PyMuPDF (fitz)` or `PyPDF2` for text-based

  Frontend: Streamlit or Flask Web App

  Evaluation: Manual ground truth + script to compare extracted fields

  Deployment: Local (Docker optional)

📂 Dataset & Evaluation

- **CV Corpus**: 10–20 diverse CVs (some with selectable text, others as scanned images)
- **Fields to Extract**: Name, Email, Phone, Education, Skills, Experience
- **Evaluation Metrics**: Field-level Precision, Recall, and F1 Score
- **Comparison**: Between 3 open-source LLMs

📤 Submission Guidelines

Candidates must submit:

- ✅ **GitHub Repo** with structured code, including:
  - A `run.sh` script to launch the app and LLM setup
  - Instructions for installing Ollama and models

- ✅ **Web App**: A local web app (Streamlit/Flask) for uploading and extracting CVs
- ✅ **Documentation**:
  - Setup instructions, model list, how extraction works
  - Accuracy comparison with evaluation results (charts or tables)
  - Mention any coding AI tool used (e.g., GitHub Copilot, ChatGPT)
- ✅ **Evaluation Report**: Include comparison of 3 models with extraction metrics

- ✅ **Video**: A 2–3 minute YouTube video (unlisted) explaining the system and showing a working demo

📥 **Submission Form**

Use the same Google Form:

https://docs.google.com/forms/d/1-btsLJyb_Wo12DVfgPxgaI_vqsb-gCyCwlS2qrN3Ar4/preview