# Project Report: VN Infra Hybrid AI Recruitment System

## 1. Abstract

The recruitment industry faces a significant challenge in processing the sheer volume of applications efficiently. Traditional Applicant Tracking Systems (ATS) rely on rigid keyword matching, often rejecting qualified candidates who use different terminology. This project, **VN Infra Hybrid**, introduces a next-generation recruitment platform that utilizes **Semantic Machine Learning** (BERT) for intelligent scoring and **Generative AI** (Llama-3/Gemini) for workflow automation. By combining local privacy-first processing with cloud-based intelligence, the system offers a balanced, scalable solution for modern HR needs.

## 2. Introduction

Hiring is a data-heavy process. A single job posting can attract hundreds of resumes, making manual review impossible. Existing automated solutions are either too simple (keyword counters) or too expensive (enterprise SaaS).

**VN Infra Hybrid** solves this by:

1. **Understanding Context:** Using Vector Embeddings to understand that "Python" and "Django" are related skills.
2. **Ensuring Privacy:** Processing resume scoring locally on the user's machine/server.
3. **Automating Communication:** Using LLMs to draft personalized emails and interview questions.
4. **Self-Improving:** Implementing a "Human-in-the-Loop" feedback system to train the model over time.

## 3. System Architecture

The system follows a **Client-Server Architecture** with a unique Hybrid AI integration:

- **The Frontend Client:** A responsive web dashboard serves as the command center. It manages candidate data in the browser's `IndexedDB`, ensuring instant access without heavy database latency.
- **The Local Inference Server:** A Python Flask application hosts the Sentence-Transformer model. It receives PDF text, converts it into 384-dimensional vectors, and computes the cosine similarity against the Job Description (JD) vector.
- **The Feedback Loop:** User corrections (e.g., changing a score from 40% to 80%) are captured and saved to a CSV dataset. This data is used to periodically re-train the model, enabling it to adapt to specific hiring preferences.

## 4. Methodology

1. **Data Preprocessing:** Resumes are parsed using `pdfplumber`. Text is cleaned using RegEx to remove special characters and stop words.
2. **Vectorization:** The `all-MiniLM-L6-v2` model converts both the Resume and JD into vector embeddings.

3. **Scoring Algorithm:**
    - *Semantic Score (40%):* Cosine similarity between vectors.
    - *Keyword Score (60%):* Substring matching of critical hard skills found in the JD.
4. **Generative Tasks:** For qualitative tasks (Interview Questions), the system constructs a prompt with the "missing skills" and sends it to the Groq API to generate context-aware questions.

## 5. Results & Discussion

- **Accuracy:** The initial model achieved a baseline accuracy of ~53% on synthetic data. After fine-tuning with 500 augmented samples, the model demonstrated improved stability, correctly identifying seniority and related tech stacks.
- **Performance:** Resume parsing and scoring takes <800ms on a standard CPU, proving the viability of local inference.
- **User Experience:** The "Voice Interview" feature successfully converts generated text to speech, allowing for a hands-free screening experience.

## 6. Conclusion & Future Scope

VN Infra Hybrid successfully demonstrates that powerful ATS tools can be built with open-source models. It bridges the gap between privacy and intelligence.

**Future Enhancements:**

- Integration with Gmail API for auto-sending drafts.
- Deployment of the backend to a GPU-accelerated container (Docker/Kubernetes).
- Expansion of the dataset to include bias-reduction samples.

## 7. References


- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
- Vaswani, A., et al. (2017). Attention Is All You Need.
- Open Source Libraries: HuggingFace, Chart.js, Flask.