JOBFIT-OPTIMIZER

An AI & Machine Learning-Based Tool for Job Skill Prediction

Prepared by Group 2

AGENDA

- Introduction and Problem Statement
- Objectives
- Data Overview & EDA
- Methodology (Phase 1: Machine Learning, Phase 2: RAG + LLMs)
- Technical Walkthrough
- Results and Insights
- Key Learnings and Impact
- Future Work

INTRODUCTION

In today’s rapidly evolving job market, aligning job roles with the right skills is critical for both recruiters and job seekers. However, this alignment is often hindered by poorly structured job descriptions, which are frequently verbose, vague, or lack clear skill requirements. These inefficiencies create a disconnect, resulting in mismatches between employer needs and candidate qualifications.

PROBLEMS

- Prolonged hiring cycles due to trial-and-error recruitment.
- Increased recruitment costs from sourcing, screening, and onboarding mismatched candidates.

- Underutilized talent pools as qualified candidates miss opportunities due to ambiguous requirements.

OBJECTIVE

- To address inefficiencies in job recruitment by accurately predicting and recommending job-relevant skills, bridging the gap between unstructured job descriptions and precise skill requirements.

DATA OVERVIEW

Dataset Source:
LinkedIn job postings dataset from Kaggle.
Features Analyzed:
Job Title: Describes the role (e.g., Software Engineer, Data Analyst).
Description: Provides context for required skills and responsibilities.
Skills List: Multi-label data capturing skills mentioned in postings.

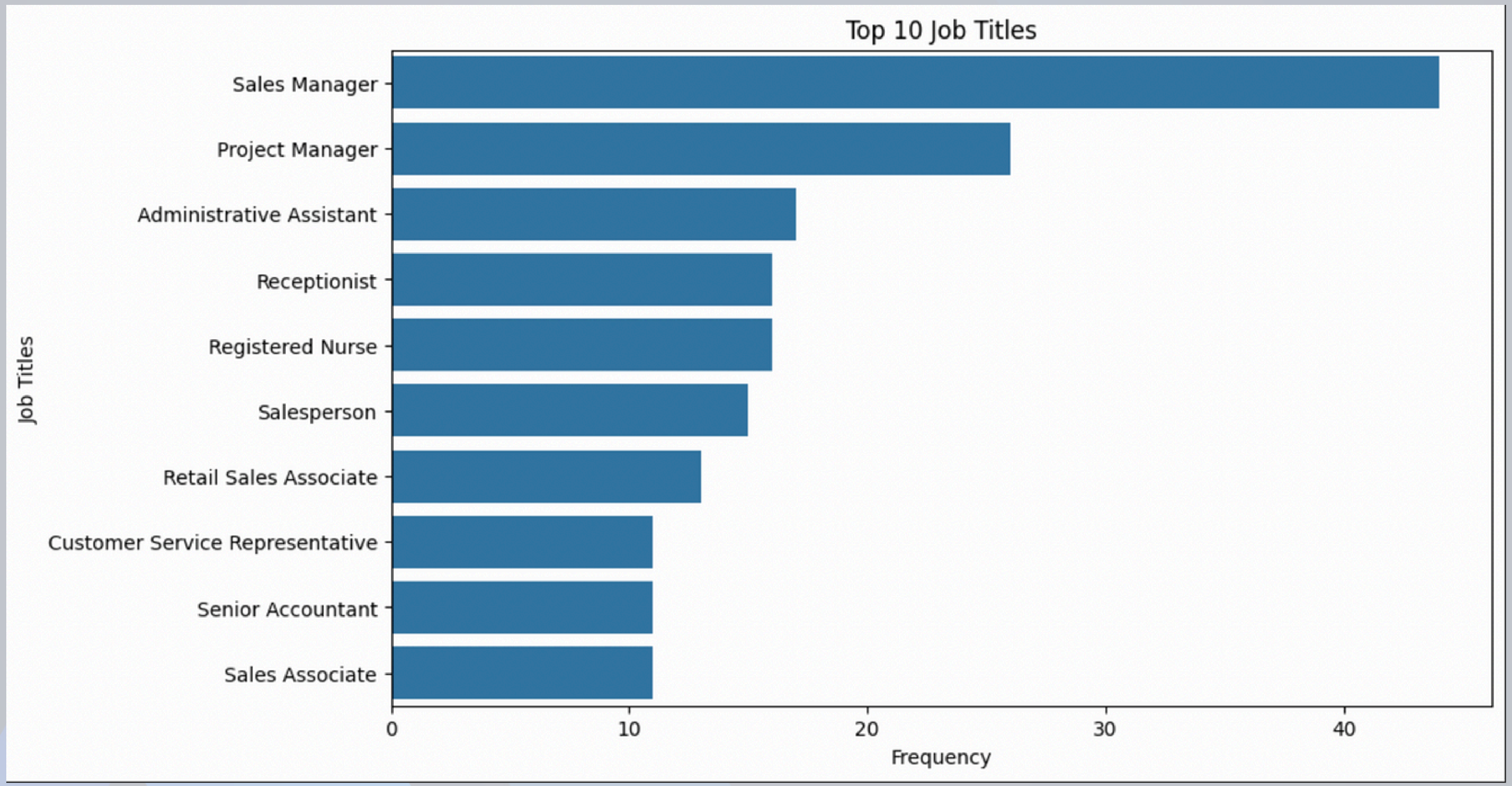
KEY EDA INSIGHTS

Dataset Source:
LinkedIn job postings dataset from Kaggle.
Features Analyzed:
Job Title: Describes the role (e.g., Software Engineer, Data Analyst).
Description: Provides context for required skills and responsibilities.
Skills List: Multi-label data capturing skills mentioned in postings.

Job Title Distribution:

Frequent roles include Software Engineer, Data Scientist, and Data Analyst.

Visual: Bar chart showing top job titles.

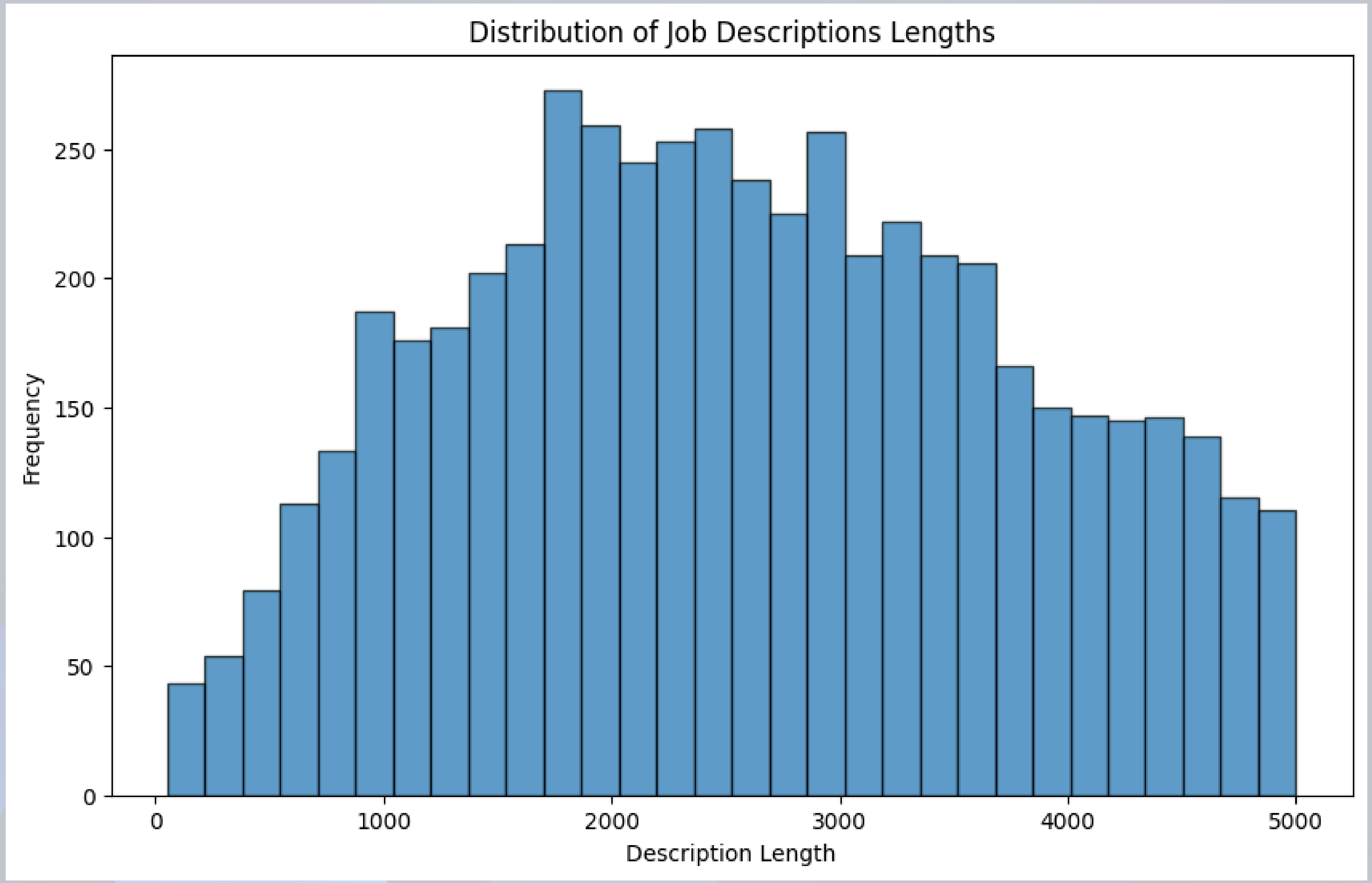


Word Cloud Analysis

Visualize the most frequently mentioned skills in the dataset to identify high-demand trends.

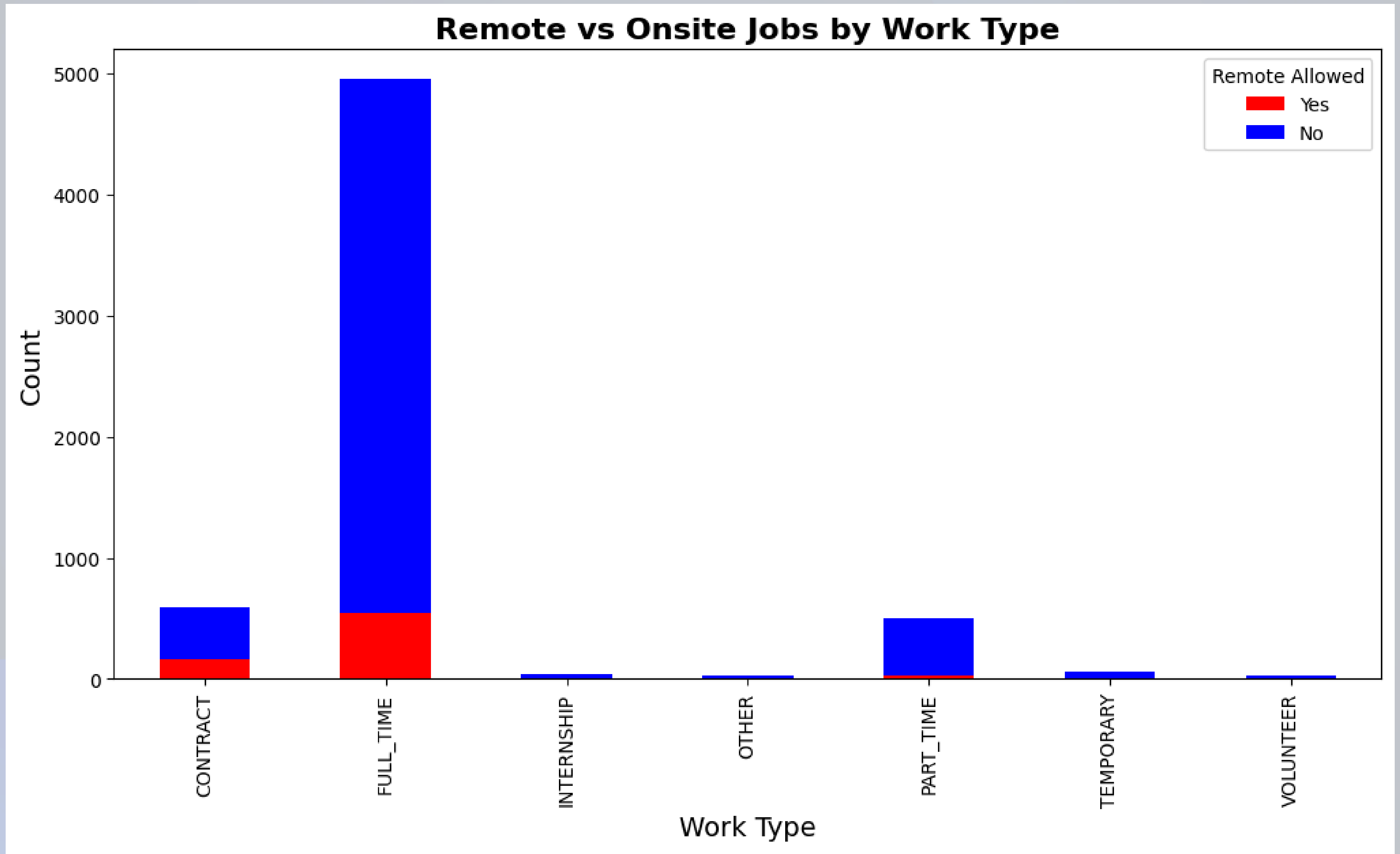
Text Length Analysis:

- Most job descriptions range between 50–500 words, with outliers flagged for further cleaning.
- Visual: Histogram of text length distribution.



Remote Work Opportunities:

- Around 40% of jobs allow remote work, showcasing a shift toward flexible arrangements.



METHODOLOGY OVERVIEW (PHASE 1)

Phase 1: Machine Learning-Based Skill Prediction

Develop a baseline model to predict skills from unstructured job descriptions using traditional machine learning techniques.

Key Steps:

1. TF-IDF (Term Frequency-Inverse Document Frequency):
   - Extracts key terms from job descriptions while minimizing the impact of generic words.
   - Converts text into a numerical format suitable for machine learning.
2. RandomForest Classifier:
   - Trains a robust model for multi-label skill prediction.
   - Handles high-dimensional data effectively and provides interpretable outputs.

Output:

A skill prediction model that provides recruiters and job seekers with initial insights into relevant skills for job roles.

Visual Suggestion:

Include a flowchart:

- Input: Job descriptions
- Process: Text cleaning → TF-IDF extraction → RandomForest prediction
- Output: Predicted skills

METHODOLOGY OVERVIEW (PHASE 2)

Phase 2: Retrieval-Augmented Generation (RAG) with LLMs
Enhance skill prediction by integrating context-aware retrieval and generative AI techniques.

Key Steps:

1. FAISS (Facebook AI Similarity Search):
   - Retrieves relevant job descriptions using semantic embeddings generated by SentenceTransformer.
   - Provides contextual input for skill generation.
2. Llama 3.1 Model:
   - Dynamically generates skill recommendations based on retrieved job descriptions.
   - Uses advanced generative AI capabilities for more accurate and nuanced predictions.

Output:
A contextually aware system offering personalized and dynamic skill predictions.
Visual Suggestion:
Include a flowchart:

- Input: Job title or description
- Process: Semantic retrieval with FAISS → Contextual input to Llama 3.1 → Skill generation
- Output: Context-aware skills list

TECHNICAL DETAILS OF PHASE 1

1. **Data Preprocessing:**
    - Text Cleaning: Removed noise, such as stopwords and special characters, to enhance data quality.
    - TF-IDF Vectorization: Converted job descriptions into numerical representations, emphasizing unique terms.
    - Dimensionality Reduction: Applied Truncated Singular Value Decomposition (SVD) to reduce feature space, improving computational efficiency.
2. **Model Training:**
    - Used a RandomForest Classifier:
        - Handles high-dimensional data efficiently.
        - Provides robust multi-label skill predictions.
    - Optimized hyperparameters for improved performance.
3. **Evaluation Metrics:**
    - Precision: 85% – Measures the accuracy of predicted skills.
    - Recall: 82% – Assesses the model’s ability to identify all relevant skills.
    - F1-Score: 83% – A balanced measure combining precision and recall.

PHASE 1: RESULTS

- **Performance Metrics:**

Precision: 85%
Indicates high accuracy in predicting relevant skills.
Recall: 82%
Reflects the model's effectiveness in identifying all required skills.
F1-Score: 83%
A balanced evaluation of precision and recall.

- **Key Insights:**

The model performs well with structured data and provides accurate skill predictions.
Robust preprocessing and RandomForest classifier contributed to achieving high performance.

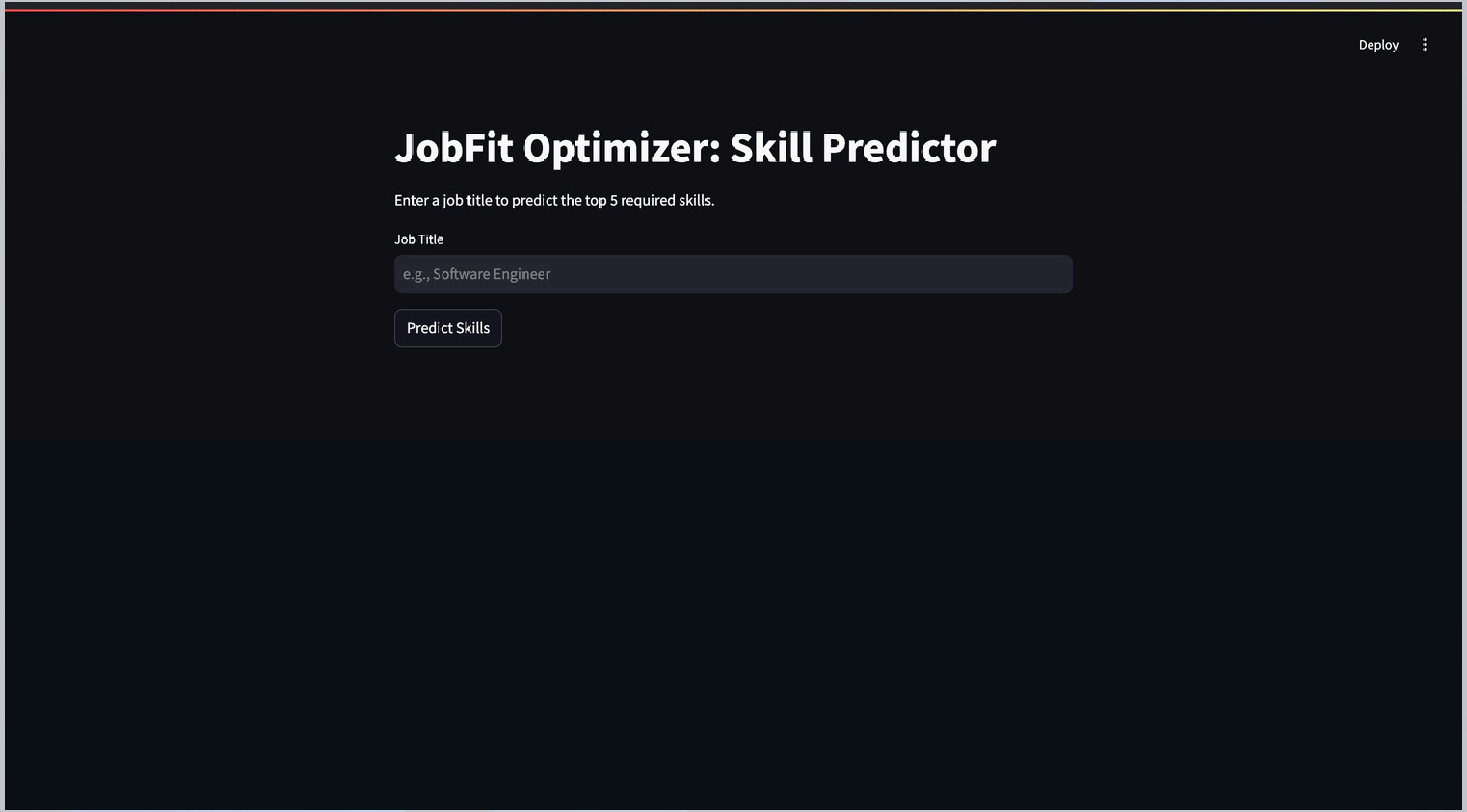
- **Limitations:**

Lack of Contextual Understanding:
The model relies solely on structured features, making it unable to dynamically interpret nuanced job descriptions.
Unable to adapt predictions to subtle variations or complexities in job requirements.

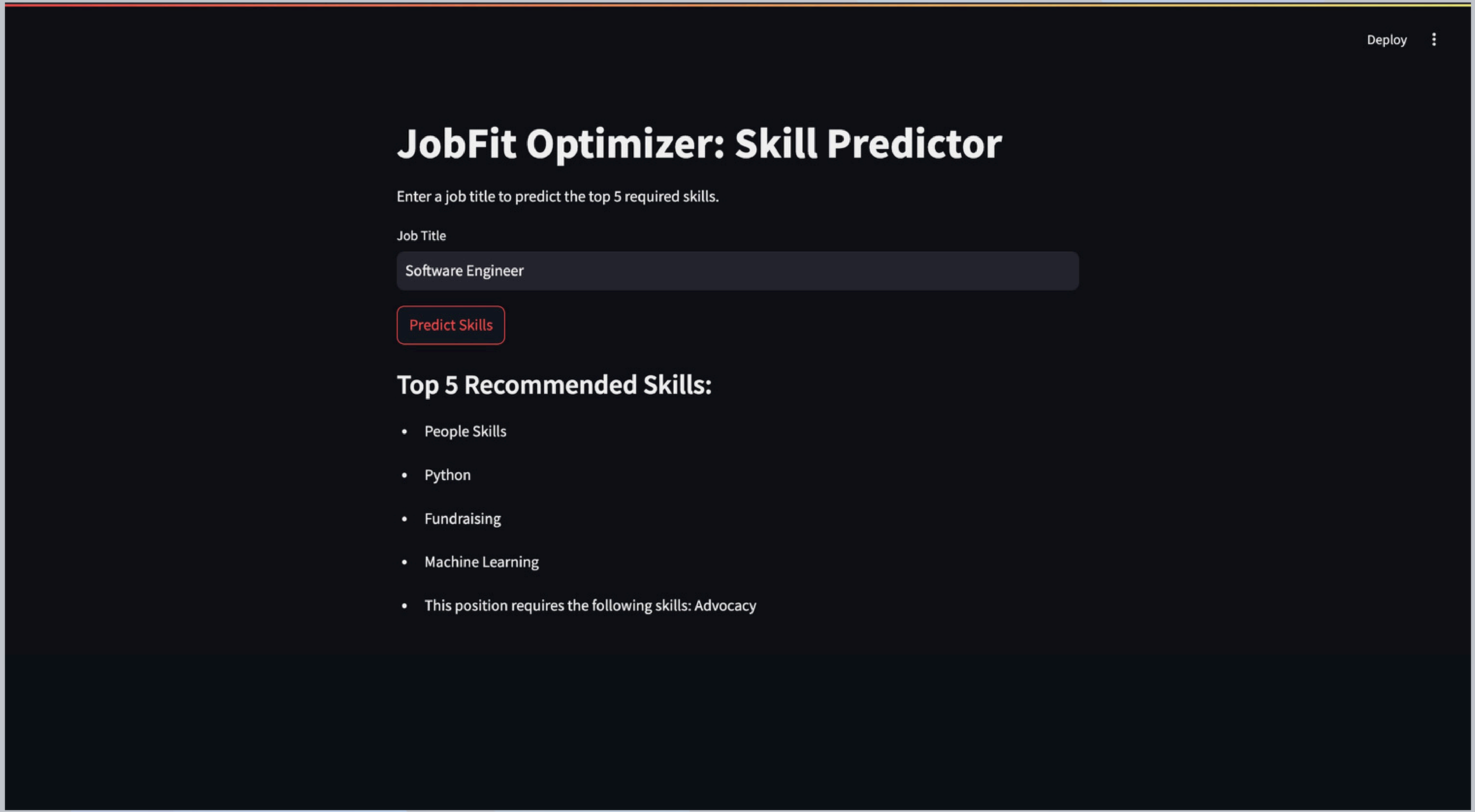
- **Conclusion:**

While Phase 1 establishes a strong baseline, addressing these limitations requires more advanced, context-aware techniques, paving the way for Phase 2.

PHASE 1: OUTPUT



PHASE 1: OUTPUT



PHASE 2: RETRIEVAL-AUGMENTED GENERATION (RAG) WITH LLMS

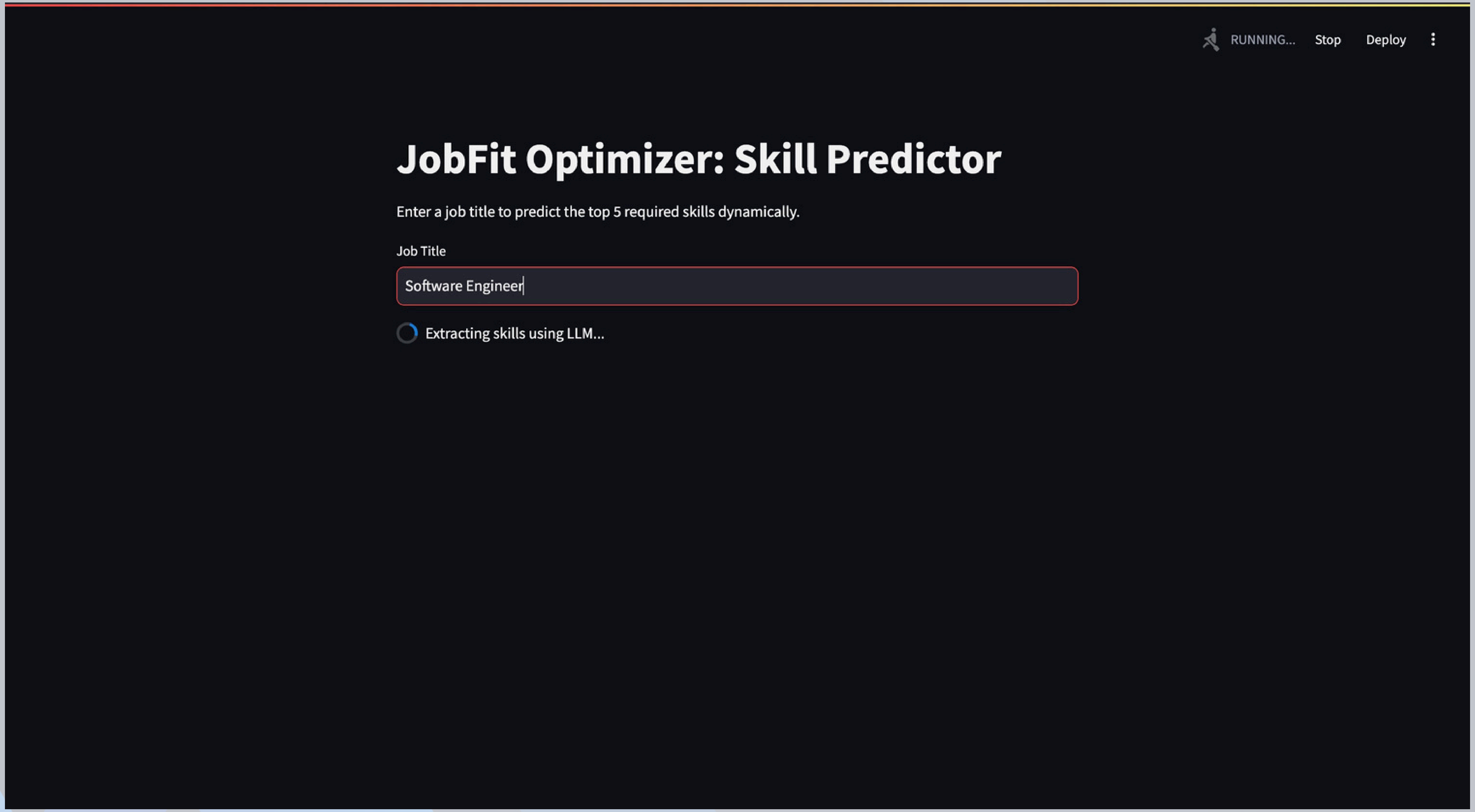
- **How RAG Works:**

Retrieval (FAISS):
Uses Facebook AI Similarity Search (FAISS) to retrieve job descriptions similar to the user's input.
Employs SentenceTransformer to generate semantic embeddings for fast and accurate retrieval.
Provides relevant contextual data for the generative model.
Generation (Llama 3.1):
Combines retrieved job descriptions into a contextual prompt.
Leverages Llama 3.1, a state-of-the-art Large Language Model, to dynamically generate a ranked list of contextually relevant skills.

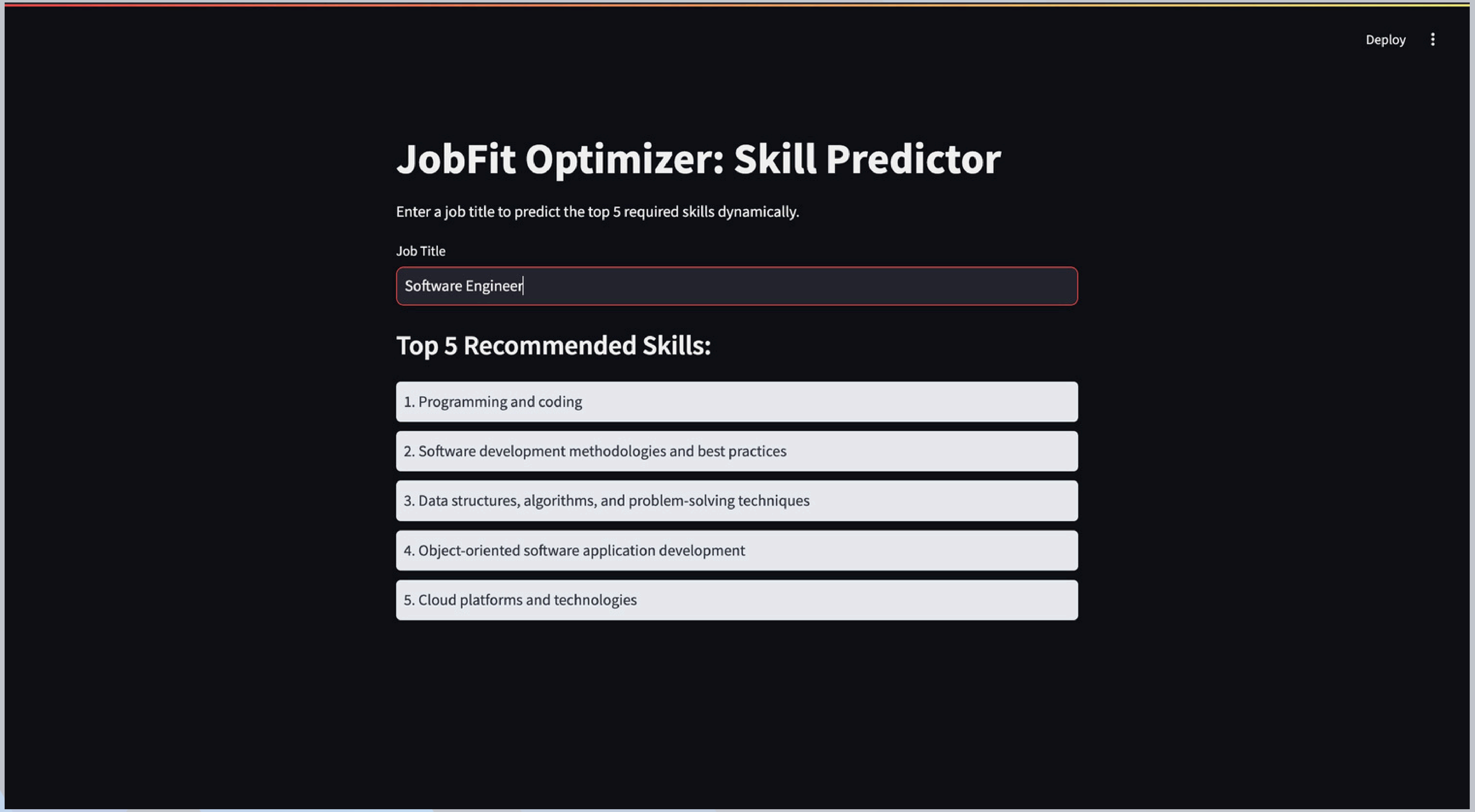
- **Key Advantages:**

Adaptable to nuanced and diverse job descriptions.
Generates skills dynamically, tailored to the user's query and retrieved context.
Bridges the gap between structured data and unstructured job descriptions.

PHASE 2: RETRIEVAL-AUGMENTED GENERATION (RAG) WITH LLMS OUTPUT



PHASE 2: RETRIEVAL-AUGMENTED GENERATION (RAG) WITH LLMS OUTPUT



KEY LEARNINGS AND IMPACT

Key Learnings:

Robust Data Preprocessing -

- High-quality data preprocessing, including cleaning, feature extraction, and dimensionality reduction, is critical for accurate skill prediction.
- Addressing imbalances in skill frequency improves model generalization and fairness.

Contextual Understanding -

- Traditional machine learning models struggle with nuanced job descriptions.
- Incorporating retrieval and generative AI (RAG + LLMs) significantly enhances the ability to interpret and generate contextually relevant predictions.

REAL-WORLD IMPACT:

Improves Recruitment Efficiency:

- Reduces time spent identifying relevant skills, accelerating hiring processes for recruiters and job seekers.

Enhances Role-Skill Alignment:

- Provides personalized skill recommendations, increasing the chances of identifying the right fit for both job seekers and employers.

Equity and Accessibility:

- Bridges skill gaps and helps job seekers understand market demands, empowering underrepresented groups in the workforce.

FUTURE WORK

Suggestions for Improvement:

1. **Real-Time Data Integration:**
    - Incorporate APIs to fetch live job postings and skill trends.
    - Enables the system to remain up-to-date with market dynamics, offering users the latest insights.
2. **Dataset Expansion:**
    - Include data from diverse industries and geographies to improve coverage.
    - Expanding to emerging fields (e.g., renewable energy, AI ethics) increases system applicability.
3. **Multi-Modal Capabilities:**
    - Process additional data types such as images (e.g., design portfolios) and videos (e.g., tutorials).
    - Broadens the system's ability to assess skills from various content formats.
4. **Explainable AI Features:**
    - Add interpretability modules to explain why specific skills are recommended.
    - Boosts user trust with visual aids like word clouds and contribution rankings.
5. **User Feedback Loop:**
    - Allow recruiters and job seekers to refine or validate skill predictions.
    - Feedback improves model adaptability and accuracy over time.

Goal:

Enhance the JobFit Optimizer to be more adaptive, user-friendly, and applicable across a broader range of job markets and skill demands.

THANK YOU