

AI-Powered Recruitment Systems: Conversational Assessment and Predictive Shortlisting

Dhanashri Patil, Kaveri Aher, Atharva Bhoite, Apurva Bhosale

Department of AI & Data Science

AISSMS Institute of Information Technology, Pune, India

patildhanashriramrao@gmail.com, kavariaher843@gmail.com

apurvabhosale15504@gmail.com, atharvabhote@gmail.com

Abstract: *The rapid adoption of Artificial Intelligence (AI) in Human Resource Management (HRM) demands a unified, objective framework to address recruitment challenges such as human bias, lengthy cycles, and inconsistent candidate evaluation. This survey details an end-to-end AI-powered recruitment solution built upon a Django Job Portal and functioning across three stages. The pipeline begins with Automated Skill Matching using NLP and semantic models for high-accuracy initial shortlisting. Candidates exceeding the relevance threshold proceed to the core innovation: the AI-Powered Voice-Based Interview. This module utilizes a cascaded Speech-to-Text (STT), Large Language Model (LLM), and Text-to-Speech (TTS) architecture to simulate dynamic, conversational assessments that adapt questions based on the candidate's technical responses. Crucially, the system employs locally-hosted LLMs, such as Ollama, for in-house inference. This architectural choice ensures superior data privacy and security by preventing sensitive candidate data from being transferred to external cloud APIs, while simultaneously achieving greater cost-efficiency for high-volume recruitment. By integrating predictive analytics for final shortlisting, this research provides a comprehensive blueprint for building scalable, unbiased, and equitable hiring ecosystems that enhance both operational efficiency and ethical accountability.*

Keywords: AI in Recruitment, Automated Screening, Voice- Based Interview, LLM-as-an-Interviewer, Predictive Hiring, Algo- rithmic Bias, Natural Language Processing

I. INTRODUCTION

The digitalization of Human Resource Management (HRM) has been significantly accelerated by the widespread adoption of Artificial Intelligence (AI), transforming traditional recruitment workflows into data-driven, adaptive, and automated processes [1], [2]. Conventional recruitment systems often face challenges such as high turnover rates, inherent human bias in evaluation, prolonged hiring cycles, and a lack of standardization in interview procedures [3], [4]. These limitations emphasize the urgent need for intelligent hiring systems capable of evaluating candidates not just based on static resumes but through dynamic, multi-dimensional assessment frameworks.

In recent years, organizations have begun leveraging AI for diverse HR functions such as resume parsing and candidate ranking. However, most existing solutions operate in isolation, handling only one stage of recruitment, such as simple resume screening or pre-scripted chatbot interviews [5]. There remains a notable research and implementation gap in developing a unified system that integrates all stages—screening, testing, interviewing, and final shortlisting—into a single, coherent pipeline. Addressing this gap, the present study introduces an AI-powered, end-to-end hiring architecture that automates the candidate journey from application submission to final recruiter recommendation [6], [7]. The foundational component of the user platform is a secure Django Job Portal implementation, ensuring robust application management [8]. The proposed system operates through three sequential AI-driven stages. The first stage, Automated Skill Matching, utilizes Natural Language Processing (NLP) techniques and semantic similarity models to evaluate the contextual fit between candidate profiles and job descriptions [9]. The second stage, AI-powered Voice-based Interview, incorporates advanced Speech-to-Text (STT) and Text-to-Speech (TTS) frameworks in conjunction



with Large Language Models (LLMs) to simulate human-like technical and HR interviews. These conversational assessments dynamically adapt questions to candidate responses, thereby improving contextual relevance and engagement [5], [6].

A key differentiator of this research is the strategic deployment of the LLM component. Unlike proprietary cloud solutions that incur unpredictable token costs and raise critical data sovereignty concerns [4], [10], our system utilizes locally-hosted LLMs (via Ollama). This architecture ensures superior data privacy by keeping sensitive candidate transcripts and evaluations within the organization's private network, simultaneously delivering significant cost-efficiency for high-volume use cases [11]–[13].

Furthermore, the paper discusses critical factors such as fairness, transparency, and explainability in AI-driven recruitment. The incorporation of bias mitigation techniques and the use of Explainable AI (XAI) models ensure ethical alignment and equitable evaluation of all applicants [14], [15]. This systematic integration not only enhances operational efficiency but also promotes inclusivity and trust in automated hiring decisions.

A. System Scope and Contribution

The proposed system operates as a dual-purpose platform serving both candidates and recruiters, thereby bridging the gap between job seekers and organizations through AI-driven automation. For candidates, it functions as a comprehensive job portal offering profile creation, skill-based matching, and real-time interview experiences. For recruiters, it serves as an intelligent Applicant Tracking System (ATS) capable of analyzing, ranking, and predicting candidate success probabilities based on multi-stage assessment results.

At its core, the system integrates a three-phase pipeline that mirrors the structure of conventional recruitment while embedding AI in every stage for consistency and fairness:

II. LITERATURE REVIEW

The application of Artificial Intelligence within recruitment and selection processes is a highly active field of study, driven by the need for greater efficiency, scale, and objectivity in talent acquisition [1], [2]. This section synthesizes research across three main thematic areas to contextualize our proposed integrated, multi-stage hiring system.

A. Automated Screening and Semantic Matching

The study by Rathore et al. [1] provides a strong foundation for the entire project, analyzing the overall impact of AI on recruitment and selection processes. This paper establishes that AI-driven automation enhances procedures, contrasting traditional time-consuming methods with automated approaches. It validates the foundational shift toward technology in the hiring lifecycle, setting the stage for more complex, multi-modal systems.

The work by Wang and Zhang [3] focuses specifically on the effectiveness of Applicant Tracking Systems (ATS), providing a crucial critique of current limitations in automated screening. Their research contrasts traditional keyword-based matching, which is often prone to bias and inefficiency, with modern AI approaches that aim to mitigate subjective judgments, thus justifying our system's core focus on high-accuracy, objective filtering.

The research conducted by Patel and Singh [16] is fundamental to the screening module, demonstrating how Natural Language Processing (NLP) is used for robust resume parsing. This paper specifically addresses the use of NLP models to accurately extract and analyze candidate skills and qualifications, which directly supports our architecture's reliance on accurately converting unstructured resume data into measurable features.

Sharma and Gupta's work [9] delves into the technical advantage of semantic matching using deep learning, a core methodology in our screening phase. This paper confirms that methods calculating vector similarity (like Cosine Similarity) significantly outperform older lexical approaches in accurately ranking candidates based on contextual relevance, providing the theoretical backing for our advanced semantic filtering algorithm.

The study by Kabade et al. [2] contributes to the understanding of early-stage interview automation, presenting a system for AI-enabled evaluation. While focusing on general automated interview frameworks, the research validates the technical viability of combining different AI components to objectively score candidates, ensuring that foundational interview modules are both functional and scalable within a hiring pipeline.



The findings by Alomari et al. [8] provide architectural justification for the application platform, detailing the need for a robust and secure web-based job portal utilizing the Django Framework. This reference supports our design choice, emphasizing that the candidate management interface must prioritize security and reliable infrastructure to handle sensitive applicant data effectively.

B. Conversational AI for Dynamic Interviewing

Allbert et al. [5] offer a detailed technical evaluation of the primary conversational architecture—the cascaded STT → LLM → TTS pipeline. Their research is critical to our system design as it analyses the effectiveness and modularity of this sequence, while also highlighting the inherent challenge of error propagation from the Speech-to-Text component, which necessitates robust transcription layers.

The paper by Kim et al. [6] introduces and validates the concept of the LLM-as-an-Interviewer, a key functional element of our conversational assessment phase. This work demonstrates that a Large Language Model can dynamically adapt its questioning based on previous responses, thereby providing a deeper, more contextually relevant evaluation than static, pre-scripted interviews.

Jabarian and Henkel's academic study [7] provides empirical evidence supporting the use of voice-based AI for high-stakes hiring decisions. Their natural field experiment confirms that AI-led voice interviews can statistically outperform traditional human screening methods in objective outcomes such as job offer rates and long-term candidate retention, justifying the implementation of our entire voice-based module.

The research by Patil et al. [17] addresses the critical feature of soft-skill assessment within voice interviews. This paper demonstrates the integration of real-time voice and emotion analysis alongside semantic analysis, allowing our system to provide objective scoring on non-verbal communication elements like confidence, composure, and tone.

The study on robust Speech-to-Text conversion by Singh and Sharma [18] addresses a significant technical vulnerability in any voice-based system. Their work focuses on enhancing STT accuracy in difficult scenarios, such as noisy environments or accented speech, which is essential for ensuring the fairness and reliability of the LLM's evaluation process in diverse settings.

Cui et al.'s survey [19] provides comprehensive coverage of Speech Language Models (SpeechLM), which represents the next frontier in conversational AI. This paper is cited to highlight the limitations of the cascaded STT-LLM-TTS architecture and to identify future work goals, specifically moving towards end-to-end models that preserve paralinguistic information lost in text-only intermediate steps.

C. Ethical Frameworks and Strategic Deployment

The paper by Soni [20] is foundational to our ethical section, analyzing the core necessity for efficiency and equity in automated hiring. This reference emphasizes that the system design must explicitly incorporate fairness metrics and bias mitigation strategies to ensure the process remains equitable and prevents AI from reproducing historical discrimination.

The work of Bose and Ray [4] focuses on the broader ethical and legal considerations of AI-based hiring systems. It highlights compliance risks, data privacy mandates, and the necessary balance between automation and legal frameworks, reinforcing the need for XAI and data governance in our multi-stage architecture.

The study by Abayomi et al. [14] validates the procedural necessity of continuous algorithmic auditing and bias detection. By analyzing case studies of automated decision systems, this research stresses that bias is an accumulated problem, requiring constant oversight and objective fairness measurements throughout the AI-driven pipeline.

The research by Al-Musawi et al. [15] advocates for the role of Explainable AI (XAI) in building trust in automated HR systems. This directly supports our objective of providing transparency, ensuring that recruiters are not merely presented with a final score but can understand which features and criteria drove the hiring recommendation.

Sahu and Singh's analysis [10] provides the economic justification for our non-cloud approach, performing a cost-benefit analysis of on-premise LLM deployment. This key reference establishes that moving away from unpredictable token-based pricing via commercial APIs results in a lower Total Cost of Ownership (TCO) for high-volume



recruitment. The research by Rodríguez and Albo's [11] provides specific technical grounding for utilizing Ollama in a privacy-focused local LLM deployment. Their work validates that hosting LLMs internally keeps sensitive data within a closed environment, guaranteeing data sovereignty and security against external breaches, which is a key selling point of our system.

The work of R. Soni and S. Kumar [12] focuses on performance evaluation of local LLMs on edge devices. This is crucial technical backing for our Ollama implementation, confirming that even resource-constrained edge systems (or local servers) can handle the real-time NLP and inference tasks required by the conversational AI modules with acceptable latency.

Finally, the comprehensive survey on LLM-based Edge Intelligence [13] establishes the cutting-edge nature of our local LLM strategic architecture. This paper summarizes the state of the art regarding security, trustworthiness, and architectural feasibility of moving large models off the cloud, providing strong context for why our solution is both novel and necessary in the current AI landscape.

III. RELATED WORK AND COMPARATIVE ANALYSIS

Artificial Intelligence in recruitment has evolved significantly over the past decade, with initial systems focusing heavily on rule-based keyword matching, which often failed to capture the contextual relevance of candidate experience [3]. With advancements in Natural Language Processing (NLP), the focus shifted toward vector-based semantic understanding of resumes and job descriptions using models like Word2Vec and BERT [16]. These models improved match accuracy but still lacked the real-time adaptability and conversational depth required for objective assessment. Modern recruitment platforms integrate multiple modalities (text, speech, and even facial recognition) to assess candidates more holistically. For example, commercial systems often leverage advanced Natural Language Understanding (NLU) and dialogue systems for candidate engagement [2]. However, these solutions face inherent limitations regarding deployment flexibility, data security, and ethical governance.

A. Critique of Existing Systems

To contextualize the novel contributions of our framework, we critique major existing AI recruitment models based on deployment strategy and core functionality:

- **Commercial Cloud LLMs** (e.g., OpenAI/GPT APIs): These systems offer high performance and scalability but are fundamentally transactional, resulting in unpredictable token-based pricing and significant concerns over data sovereignty and security [4], [10]. Transferring sensitive candidate data to external cloud environments poses a major governance risk.
- **Proprietary Behavioral Systems** (e.g., HireVue/Other Systems): These platforms often rely on opaque computer vision or emotion analysis models that assess soft skills based on facial cues or body language. These "black box" approaches introduce significant risks of algorithmic bias and lack the necessary Explainable AI (XAI) components to justify decisions ethically [14], [15].
- **Academic/Open-Source Prototypes**: While research has validated the LLM-as-an-Interviewer concept [6], many prototypes remain isolated to single stages (e.g., text-only chatbots) or rely on cascaded architectures that are prone to transcription error propagation [5], [18].

B. Comparative Analysis Table

C. Novelty and Contribution of the Proposed System

The primary novelty of our proposed framework lies in its strategic fusion of advanced AI functionality with enterprise-critical deployment constraints.

- **TCO and Privacy Optimization (Ollama)**: By deploying LLMs locally via Ollama, we resolve the dual challenges of high, recurring token costs and data security risks inherent in cloud solutions. This shift provides a predictable Total Cost of Ownership (TCO) and ensures maximum data governance [10], [11], [13].
- **Integrated End-to-End Voice Assessment**: We unite semantic screening [9], the conversational LLM-as-an-



Interviewer [6], and objective soft-skill analysis [17] into a single, seamless voice-driven pipeline, significantly advancing the realism and utility beyond conventional chatbot or static systems.

TABLE I: COMPARATIVE ANALYSIS OF AI RECRUITMENT SYSTEMS

Feature / Criterion	Commercial Cloud LLMs	Proprietary Behavioral ATS	Proposed System (Ollama-Django)
Core Deployment Model	External API (AWS, Azure)	Private Cloud / SaaS	Local/On-Premise Inference
LLM Access Method	Token-based consumption	Proprietary/Not Applicable	Ollama/OpenChat (Self-hosted) [11], [12]
Data Privacy and Control	Low (Data leaves network)	Medium (Vendor-controlled)	High (Data Sovereignty Guaranteed) [11]
Pricing Model	High/Unpredictable (Pertoken)	Subscription (Fixed/Tiered)	Low/Predictable (Fixed Infrastructure TCO) [10]
Interview Modality	Text Chatbot or Pre-recorded video	Video/Facial Analysis	Dynamic Two-Way Voice/Speech [7]
Transparency / XAI	Minimal or API-based rationale	Low ("Black Box" Models)	High (XAI Scorecards Integrated) [15]
Bias Risk Focus	Embedded in training data	Embedded in proprietary models (CV/Facial)	Continuous Auditing & Semantic Focus [14], [20]

- **Ethical by Design: The reliance on semantic NLP:** [16] over potentially biased visual cues, combined with mandatory XAI dashboards and continuous algorithmic auditing [14], [15], ensures the system is built with accountability and fairness at its core. This integrated, privacy-focused architecture thus represents a significant contribution to building efficient, trustworthy, and scalable intelligent hiring ecosystems.

IV. SYSTEM ARCHITECTURE

The overall system architecture, illustrated in Figure 1, is structured as a multi-stage pipeline designed for secure data handling and efficient, end-to-end automation. This pipeline connects the external user interface with internal processing and local AI inference modules.

The architecture is composed of three primary logical layers: the User Interface Layer, the Processing Layer, and the core AI Interview Layer, centered around a privacy-focused local LLM deployment.

A. User Interface Layer (Django Web Portal)

This layer serves as the secure entry point and application management platform, built using the Django framework [8]. It handles all initial interactions:

Candidates utilize the portal to submit applications and begin the hiring process.

Recruiters define and Post Job Requirements, managing candidate flow and accessing final shortlisting reports.

B. Processing Layer (Resume Analyzer)

The Resume Analyzer initiates Phase I of the automation. Upon receiving a new application, this module immediately processes the resume data:

NLP Skill Matching: Extracts structured data (skills, experience, education) using NLP techniques [16].

Eligibility Scoring: Computes a semantic similarity score against the job description [9]. Only candidates meeting the threshold proceed to the next phase, drastically reducing recruiter workload.



C. AI Interview Layer (AI Interview System)

This module orchestrates the voice-based conversational assessments (Phase II), integrating cascaded speech technologies with a locally-hosted LLM:

- **Local LLM Deployment (Ollama - openchat):** The intelligence engine uses an LLM deployed locally via Ollama [11], [12]. This deployment ensures all sensitive candidate transcripts and inference activities remain entirely within the secure network, guaranteeing data sovereignty and predictable operational costs [10], [13].
- **Speech Processing Pipeline:** A cascaded architecture handles two-way verbal communication: Speech Recognition (STT) transcribes candidate audio to text for LLM input [18], and Text-to-Speech (TTS) converts the LLM’s dynamic questions into natural voice output [5].
- **Evaluation & Feedback:** The LLM performs real-time semantic analysis and behavioral scoring [17] to evaluate responses and dynamically generate appropriate follow-up questions [6].

All resulting data is fed into the central Database for use in Predictive Shortlisting (Phase III).

V. CONCLUSION AND FUTURE WORK

The AI-driven, multi-stage recruitment framework presented in this paper successfully streamlines the hiring pipeline—from automated semantic screening to interactive voice-based evaluation and predictive final shortlisting. By combining advanced Natural Language Processing (NLP) [16], strategic local Large Language Model (LLM) deployment [11], and adaptive speech technologies [7], the framework results in a more objective and data-centric approach to candidate assessment.

The novelty of this architecture lies in its commitment to operational efficiency and ethical governance. Utilizing locally-hosted LLMs (via Ollama) resolves critical industry concerns related to external data sovereignty and unpredictable token costs, offering a secure and financially sustainable solution for high-volume users [10], [13]. Furthermore, the framework

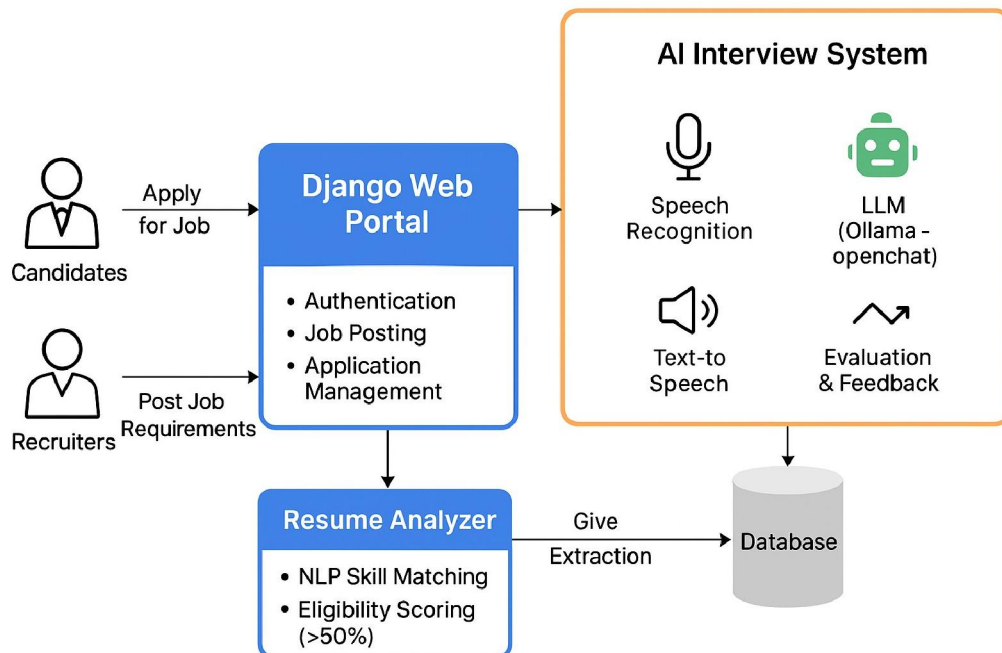


Fig. 1. Proposed AI-Powered Recruitment System Architecture

integrates fairness-aware algorithms and Explainable AI (XAI) techniques, ensuring accountability and trust in AI-assisted HR decisions [15], [20].



A. Future Work

Future enhancements will focus on extending the framework's capabilities and robustness:

- **Unified Speech-Language Models:** Transitioning from the cascaded STT-LLM-TTS architecture to emerging end-to-end SpeechLM architectures that jointly process audio and text, ensuring richer paralinguistic and emotional understanding [19].
- **Speech and Accent Robustness:** Improving the adaptability of STT systems to handle multilingual and accented speech patterns through fine-tuned domain-specific datasets [18].
- **Multimodal Analysis:** Extending candidate evaluation beyond voice to include real-time facial expression and posture recognition using computer vision models to enhance soft-skill assessment [17].
- **Ethical Governance:** Implementing automated fairness auditing, explainable scoring dashboards, and regular retraining protocols to continuously minimize bias and algorithmic drift over time [14].
- **Performance Evaluation:** Developing benchmarking protocols to rigorously assess model accuracy, system latency, and user satisfaction under various concurrent interview conditions [12].

In summary, this framework represents a significant step toward the future of intelligent hiring—where automation coexists with ethical human oversight, ensuring that recruitment remains both efficient and equitable across all stages.

REFERENCES

- [1]D. S. P. S. Rathore, "The impact of ai on recruitment and selection processes: Analysing the role of ai in automating and enhancing recruitment and selection procedures," *International Journal For Global Academic & Scientific Research (IJGASR)*, vol. 2, pp. 78–93, 2023.
- [2]V. Kabade, G. Patil, S. Godse, A. Jain, and M. Kumbharden, "Ai-enabled automated interview evaluation system," *IJIRMP*, vol. 13, pp. 1–12, 2025.
- [3]X. Wang and Y. Zhang, "Ai-powered applicant tracking systems: Efficiency and bias considerations," *International Journal of AI in Hiring*, vol. 8, pp. 112–125, 2019.
- [4]M. Bose and S. Ray, "Ethical and legal considerations in ai-based hiring systems," *International Journal of Business Ethics and AI*, vol. 7, pp. 67–83, 2022.
- [5]R. Allbert, A. Ansari, N. Yazdani, A. Mahajan, S. S. Mousavi, and A. Afsharrad, "Evaluating speech-to-text llm text-to-speech combinations for ai interview systems," *arXiv preprint arXiv:2507.16835*, 2025.
- [6]E. Kim, J. Suk, S. Kim, N. Muennighoff, D. Kim, and A. Oh, "Llm-as- an-interviewer: Beyond static testing through dynamic llm evaluation," *arXiv preprint arXiv:2412.10424*, 2025.
- [7]B. Jabarian and L. Henkel, "Voice ai in firms: A natural field experiment on automated job interviews," *Academic Study*, 2025.
- [8]A. Z. Alomari, A. K. Al-Tarawneh, and S. M. Al-Shara, "A robust and secure web-based job portal system using django framework," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 1–7, 2019.
- [9]A. Sharma and M. Gupta, "Deep learning-based semantic matching for candidate ranking in e-recruitment," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 2501– 2508.
- [10]P. Sahu and R. Singh, "A cost-benefit analysis of on-premise large language model deployment: Breaking even with commercial llm services," *arXiv preprint arXiv:2509.18101*, 2025.
- [11]A. Rodríguez and A. Albo's, "Privacy-focused llm for local data processing: Implementing ollama and rag to securely query personal files in closed environments," *O2 Repositori UOC*, 2025.
- [12]R. Soni and S. Kumar, "Performance evaluation of local llms on edge devices for real-time nlp applications," *ResearchGate*, 2024.



- [13]V. Authors, “Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness,” IEEE Open Journal of the Communications Society, 2024.
- [14]M. K. e. a. Abayomi, “Auditing algorithmic bias in automated decision systems: A case study in hiring,” MDPI, 2025.
- [15]B. M. Y. e. a. Al-Musawi, “The role of explainable ai (xai) in building trust in automated hr systems,” IEEE Access, vol. 8, pp. 150 493–150 503, 2020.
- [16]R. Patel and K. Singh, “Natural language processing for resume analysis in automated hiring systems,” IEEE Transactions on AI and Recruitment, vol. 12, pp. 78–91, 2021.
- [17]S. Patil, S. Bothara, T. Babar, and R. Kine, “Ai powered mock interview system with real-time voice and emotion analysis,” 2025 IJNRD, vol. 10, pp. 1–7, 2025.
- [18]R. Singh and P. Sharma, “Robust speech-to-text conversion in noisy environments for conversational ai,” IEEE Transactions on Speech Processing, 2021.
- [19]W. e. a. Cui, “Recent advances in speech language models: A survey,” arXiv preprint arXiv:2410.03751, vol. 14, pp. 1–17, 2025.
- [20]D. V. Soni, “Ai in job matching and recruitment: Analyzing the efficiency and equity of automated hiring processes,” 2024.

