

Advancing Recruitment: Fair and Efficient Resume Screening with LLMs

Liliana Novais^{1,2,3,4} [0000-0002-2347-6779], Vitor Rocio^{1,4} [0000-0002-3314-898X], Paulo Moura Oliveira^{4,6} [0000-0003-4283-1243] and Arnaldo Santos^{1,4,5} [0000-0001-5139-6728]

¹ Universidade Aberta, Lisboa, Portugal

² Instituto Superior de Engenharia do Porto Porto, Portugal

³ SEG Automotive Portugal, Vila do Conde, Portugal

⁴ INESC TEC, Porto, Portugal

⁵ Digimedia, Aveiro, Portugal

⁶ Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
2302227@estudante.uab.pt

Abstract. Traditional talent acquisition encounters significant sustainability problems marked by increasing application quantities and the intrinsic dangers of subjective human bias. This article, which is still being worked on, uses the Design Science Research (DSR) technique to create CurAIEval, a new IT tool for automated, fairness-aware resume screening. Based on a rigorous examination of the literature, the study finds a fundamental gap between technical optimization and ethical governance. To close this gap, the suggested architecture combines Large Language Models (LLMs) for semantic analysis with a required "Fairness-by-Design" auditing engine. The system uses a Human-in-the-Loop (HITL) protocol to automate parsing of large amounts of data while still letting people make the ultimate judgments. Expected outcomes encompass the creation of a workable prototype that substantially alleviates administrative burdens and the validation of a Balanced KPI Framework. This approach is meant to quantify both operational efficiency (like time-to-hire) and algorithmic fairness (like disparate impact) at the same time. It gives a validated technique to automate hiring in a responsible and fair way.

Keywords: LLM-Based Recruitment, Algorithmic Fairness, Resume Classification, Design Science Research.

1 Introduction

Strategic management of human capital is a key part of an organization's performance and a major source of its competitive edge [6]. Talent acquisition, or the capacity to find, attract, and hire talented people, is at the heart of this. But the first step of screening and choosing candidates is still a major problem in modern Human Resource Management (HRM). This process is sometimes very time-consuming, expensive, and manual in many companies, which means that highly skilled HR professionals must spend too much time on routine duties instead of strategic ones[8].

The digitization of job platforms and the trend to remote work after the pandemic have made this problem worse by making talent pools more global and increasing the number of applications by a huge amount. For example, SEG Automotive Portugal gets over 170 applications for each job opening, which means that it can take up to two months to choose one [15]. These kinds of delays cost a lot of money and make it more likely that you'll lose top applicants to competitors that can move faster [5].

In addition to being inefficient, manual screening also has big legal and moral risks. The procedure is quite open to unconscious bias from people, and recruiters who are under a lot of pressure may use cognitive shortcuts, which can lead to unfair results based on aspects that don't have anything to do with merit [19]. Recent progress in Artificial Intelligence (AI), particularly Large Language Models (LLMs) developed from transformer architectures [22], provides strong tools for semantic analysis and automated screening. However, the uncritical adoption of these technologies poses a "dual-use dilemma": while they provide scalability, they also risk automating and exacerbating historical biases if not meticulously regulated [13].

This study used the Design Science Research (DSR) approach [10, 17] to introduce CurAIEval, a socio-technical IT artifact, in response to the identified challenge. This approach is meant to improve hiring by using automation to make things more efficient and built-in fairness rules to make sure everyone is treated fairly.

This paper discusses the first steps of the DSR cycle, which include finding the problem and making a rough architectural plan for the CurAIEval artifact. The proposed design is based on a thorough systematic examination of the literature, although it is still a work-in-progress conceptual model. The next steps in the DSR process will be to technically implement the prototype and show it in a real-life high-volume recruitment setting. In the end, the research will end with a thorough empirical evaluation utilizing a "Balanced KPI Framework" to see how well the system can cut down on time-to-hire while also limiting disparate impact [12]. This will show that the artifact is useful in real-world situations.

2 Research Methodology

This article presents a work-in-progress whose primary objective is to design, develop, and validate a novel IT artifact to solve a practical organizational problem. Therefore, the Design Science Research (DSR) methodology was adopted [17]. DSR is uniquely suited for this work, as it focuses on creating and evaluating innovative artifacts that address real-world challenges while contributing to the theoretical knowledge base [10].

The Main Research Question (MRQ) is, '*How can contemporary AI systems enhance recruitment processes by simultaneously addressing efficiency constraints, evaluator subjectivity, HR workload, and unconscious biases while ensuring equitable selection of optimally qualified candidates?*'

To perform a systematic literature review, the MRQ is operationalized through four specific Research Questions (RQs) that the artifact must address:

- RQ1 (Frameworks): What implementation frameworks enable effective deployment of LLMs and chatbots for automated resume screening while maintaining alignment with organizational-specific hiring criteria?
- RQ2 (Balancing): Which automation strategies (e.g., scoring thresholds, human-in-the-loop protocols) most effectively balance HR workload reduction with candidate selection quality?
- RQ3 (Ethical Design): What methodological approaches best identify and mitigate unconscious biases in AI-assisted recruitment to promote diversity and equitable candidate evaluation?
- RQ4 (Metrics): Which quantitative metrics (KPIs) most accurately assess both procedural efficiency and selection fairness in AI-enhanced recruitment systems?

3 Research Contributions

This study provides unique insights at the convergence of Human Resource Management (HRM) and Artificial Intelligence (AI). Utilizing the Design Science Research (DSR) methodology [10, 17], these contributions are classified into completed contributions—originating from the problem explication and the systematic literature review—and anticipated contributions—to be achieved through the design, demonstration, and evaluation of the proposed artifact.

3.1. Completed Contributions: Foundational and Analytical

The preliminary research phase yielded three key analytical outcomes. First, the SLR identified a critical "Structural Gap" between technical optimization frameworks focused on efficiency [18] and ethical governance models centered on disability justice [14, 21]. Second, the study synthesized a scholarly consensus for Human-in-the-Loop (HITL) protocols, restricting AI to high-volume parsing while retaining human decision authority [13, 23]. Finally, we defined "Fairness-by-Design" architectural principles, including pre-processing data sanitization [1] and the integration of allocational harm metrics like RABBI directly into scoring pipelines[9].

3.2. Expected Contributions: Practical and Methodological

The subsequent DSR phases will deliver three tangible contributions:

- **The CurAIEval Artifact:** A functional prototype implementing "Fairness-by-Design" principles. Unlike theoretical models, it technically decouples fairness audits from semantic matching, offering a scalable framework for bias-aware screening that resolves manual scaling challenges [2].
- **Verified "Balanced KPI" Framework:** Addresses the deficiency in evaluation metrics [12] by integrating operational efficiency (e.g., time-to-hire) with Selection Fairness (e.g., disparate impact). This dashboard enables practitioners to assess the long-term effects of automation, as suggested by [11].

4.2. The "Structural Gap" and Its Causes

The synthesis of these clusters unveiled a significant bifurcation, termed the "Structural Gap." The research landscape is distinctly polarized:

1. **Technical Optimization (The "Green" Cluster):** Research concentrated on mathematical programming and predictive analytics to enhance efficiency and minimize costs [18].
2. **Ethical Governance (The "Red" Cluster):** Research concentrated on normative ethics, disability justice, and the sociological implications of automation (Tilmes, 2022; Mujtaba, 2019)[14, 21].

This analysis indicates that this disparity arises from underlying disciplinary differences. Computer science research frequently regards justice as a quantitative restriction to be fulfilled within a cost function, while social science research challenges the inherent validity of assessing human potential, sometimes failing to provide technical alternatives. As a result, solutions are often designed to be either efficient or equitable, but seldom at the same time.

4.3. The Imperative for a Novel Artifact vs. Adaptation

A crucial inquiry emerges in Design Science Research: Why create a novel artifact (CurAIEval) instead of modifying pre-existing commercial technologies (e.g., LinkedIn Recruiter, standard ATS)? The literature presents three compelling arguments against adaptation:

- **The "Black Box" Problem:** Commercial products that are already on the market are "black boxes" that only work with certain software. Their algorithms are hard to understand, so it's hard to check their decision-making or see if they are making previous biases worse [24]. For a serious academic study, there needs to be a "white box" architecture that lets you see and change how efficiency (LLM embeddings) and ethics (fairness restrictions) work together.
- **Reactive vs. Proactive Design:** Many commercial solutions see fairness as something to fix after the fact or as a UI feature. The research, on the other hand, says that "Fairness-by-Design" is needed to effectively reduce bias. This means putting protections deep into the data input and vectorization layers [1]. This basic architectural need can't be added to old systems that were only developed for speed.
- **The Metrics Deficit:** Right now, most systems are just good at "Time-to-Hire." There is a clear shortage of solutions that natively incorporate balanced dashboards that show both "Allocational Harm" and processing speed[9]. Making CurAIEval lets you test these new, balanced measures in a safe setting.

4.4. Robust Fairness: Beyond Legal Compliance

Finally, the evaluation requires a more nuanced definition of "Fairness" than what is already offered in conventional hiring instruments. Early methods used "Fairness

through Unawareness," which meant getting rid of protected class labels like "gender." This didn't work because of proxy bias, which happens when models learn to discriminate based on features that are related, like zip codes, language patterns, or university affiliations [4].

The CurAIEval artifact is thus required by the necessity to progress beyond just legal compliance to Algorithmic Equity. This means dealing with Allocational Harm, which is when certain populations are systematically left out of economic opportunities. The literature necessitates systems that implement active mitigation strategies, including counterfactual testing (evaluating how a resume would be ranked if the gender were altered) and intersectional analysis (ensuring equity for subgroups, such as Black women, rather than merely broad categories)[3]. Because "off-the-shelf" LLMs have built-in biases, the only way to safely deploy them is to make a custom artifact with a "Fairness Engine."

5 Proposed Artifact: *CurAIEval* Objectives

The CurAIEval system is the main IT tool for this project. It was made to turn the theoretical ideas found in the literature review into a working solution. Five interconnected goals set in the problem-framing phase influence its development:

1. Bias Mitigation (RQ3): Fight implicit bias by making algorithms that put job-related skills first, which will lead to fair hiring.
2. Process Optimization (RQ1/RQ2): Use LLMs for semantic matching to improve prediction validity beyond just using keywords
3. Strategic Repositioning (RQ2): Automate low-value administrative tasks so that HR professionals can focus on high-value interactions.
4. Performance Transparency (RQ4): Set up balanced Key Performance Indicators (KPIs) that look at both how well the process works (like time to hire) and how fair it is (like disparate impact).
5. Ethical Compliance: Make sure that algorithms follow the law so that the system can be audited and defended.

5.1. Artifact Architecture and Design

CurAIEval is designed as a modular socio-technical system that is run by "Fairness-by-Design" to make these goals happen. Instead of using retroactive ethical patches, fairness constraints are built into the architecture itself. We used Unified Modeling Language (UML) to model the system and show three important points of view.

5.1.1. Functional View: Operationalizing Human Sovereignty

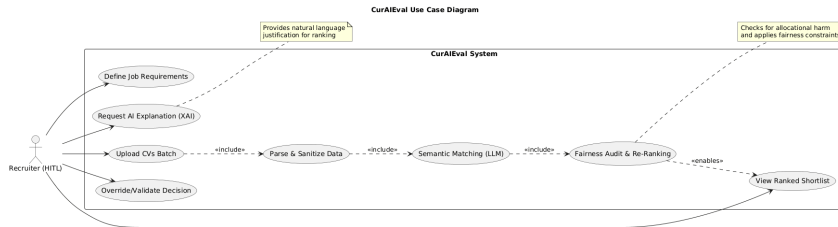


Fig. 2. Use Case Diagram for CurAIEval

The functional design (Fig. 2) rejects full automation in favor of a Human-in-the-Loop (HITL) protocol, aligning with the consensus that AI lacks the nuance required for final hiring decisions. The recruiter initiates automated tasks but retains sole authority over validation. Crucially, the "Fairness Audit" is modeled as a mandatory <<include>> relationship, preventing the bypass of bias checks for speed. Additionally, an "Explainability Request" use case addresses "black box" trust issues by providing natural language justifications for scoring.

5.1.2. Structural View: Decoupling Ethics from Intelligence

The Component Diagram (Fig. 3) talks about scalability by architecturally separating the Matching Service from the LLM API. This means that backend model modifications (such as going from GPT-3.5 to LLaMA) may be made without having to change the whole system. Most importantly, the Fairness Engine is set apart as an independent auditor. This engine maintains mathematical fairness restrictions (such as demographic parity) on the output, no matter how the semantic model behaves. This is a hard stop against allocational harm, as suggested by Delecraz et al. and Zhou[7, 25].

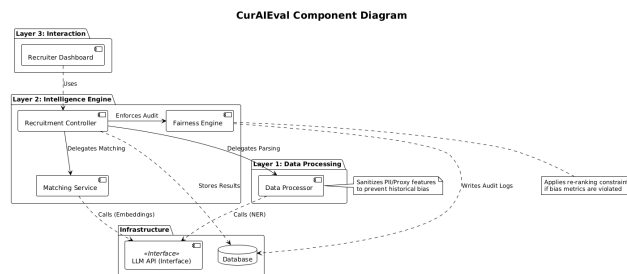


Fig. 3. Component Diagram for CurAIEval

5.1.3. Behavioral View: The Pre-Emptive Fairness Loop

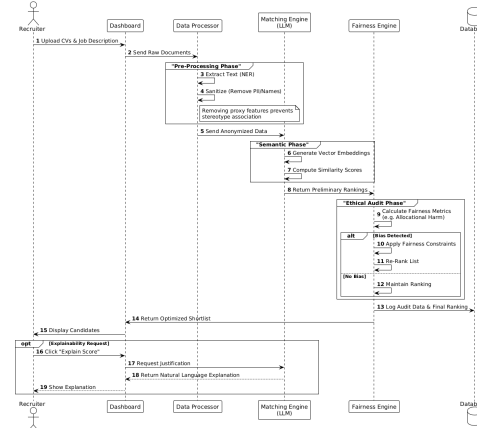


Fig. 4. Sequence Diagram for CurAIEval

The Sequence Diagram (Fig. 4) shows how data flows to get rid of the risk of "proxy bias." Before semantic embedding, a Sanitization stage removes PII (names, gender markers) to keep the LLM from encoding preconceptions. The system also uses the re-ranking logic that Gan et al. suggested. If the Ethical Audit Phase finds that a protected group is being unfairly left out, the Fairness Engine automatically starts re-ranking algorithms. This makes sure that efficiency never comes at the expense of fairness, which is what balanced metrics need.

6 Discussion and Conclusion

This study pinpoints a significant operational and ethical impediment in contemporary Human Resource Management (HRM). As established, traditional manual recruiting is increasingly unsustainable due to the globalization of talent pools and the sheer number of digital applications, resulting in significant administrative expenses and lengthy timescales [5]. At the same time, manual screening is still legally risky and open to unconscious human biases that undermine the fairness of the selection process [8]. This study, using a comprehensive Systematic Literature Review (SLR), identified a notable "structural gap" in the current academic landscape: a division between research solely concentrating on technical optimization [18] and research addressing ethical governance [21]. There are not many alternatives right now that are designed to be both operationally efficient and verifiably fair.

This work introduces CurAIEval, an IT artifact created through the Design Science Research approach [10], to address this gap. CurAIEval is not a "black box" automation tool; instead, it is thought of as a socio-technical system that follows a "Fairness-by-Design" philosophy. The system's goal is to automate high-volume parsing and semantic matching while tightly protecting human sovereignty for ultimate, context-aware decision-making by following the academic agreement on Human-in-the-Loop (HITL) protocols[13, 23]. The suggested architecture radically changes the role of the HR

professional from an administrative data processor to a strategic partner. This lowers legal risks and makes the hiring process more accurate[20].

Preliminary Design and Limitations: It is vital to note that the architectural design shown in this work is still in the early stages. The conceptual model does a good job of separating the "Fairness Engine" from the "Semantic Matcher" such that efficiency doesn't take precedence over fairness. However, the artifact has not yet been fully coded or put into use in a real setting. The current design is based on theoretical validation from the SLR instead of real-world data from field testing. As a result, the proposed algorithmic interventions, such as preprocessing data sanitization and postprocessing re-ranking, are still only theoretical ideas that need to be tested against real-world data noise and organizational complexity.

Work in the Future: The next steps in the Design Science Research cycle, Demonstration and Evaluation, will shape the direction of this research. The next step will be to work on the technical side of the CurAIEval prototype, going from architectural plans to a working piece of software that can work with current HR information systems. We will use this prototype in the real-world setting of our industrial partner, SEG Automotive Portugal, to test high-volume screening scenarios.

The most important next step is to use the proposed Balanced KPI Dashboard to test the artifact in real life. Future study will meticulously evaluate the system's effectiveness not just against efficiency criteria, such as time-to-hire and cost reduction, but also against fairness measurements, including the Disparate Impact Ratio and Allocational Harm indices [9]. Future studies will ascertain whether these automated outcomes effectively diminish the "metrics gap" identified by [12] by contrasting them with a manual baseline, thereby establishing a validated framework for the responsible and equitable implementation of large language models in recruitment.

References

1. Abdelhalim, E.: A framework of diversity, equity, and inclusion safeguards for chatbots. 67, 5, (2024). <https://doi.org/10.1016/j.bushor.2024.03.003>.
2. Alzyoud, A.A.Y.: Navigating the Future: The Role of Artificial Intelligence in Shaping Recruitment Practices. 2024 ASU Int. Conf. Emerg. Technol. Sustain. Intell. Syst. ICETISIS. (2024). <https://doi.org/10.1109/ICETISIS61505.2024.10459627>.
3. Armstrong, L. et al.: The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring - Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. Proc. 4th ACM Conf. Equity Access Algorithms Mech. Optim. (2024). <https://doi.org/10.1145/3689904.3694699>.
4. Barriuso, M.I.: Unveiling Gender Biases in Recruitment: A Natural Language Processing Approach. J. Glob. Econ. Manag. Bus. Res. (2024). <https://doi.org/10.56557/jgembr/2024/v16i18634>.
5. Cappelli, P.: The practical utility of the latest research on the time to hire. Hum. Resour. Manag. J. 19, 4, 385–397 (2009).
6. Cascio, W.F.: Managing Human Resources: Productivity, Quality of Work Life, Profits. McGraw-Hill Education (2019).

7. Delecraz, S. et al.: Making Recruitment More Inclusive: Unfairness Monitoring With A Job Matching Machine-Learning Algorithm - 2022 IEEE/ACM INTERNATIONAL WORKSHOP ON EQUITABLE DATA & TECHNOLOGY (FAIRWARE 2022). 2022 IEEE/ACM Int. Workshop EQUITABLE DATA Technol. FAIRWARE 2022. 34–41 (2022). <https://doi.org/10.1145/3524491.3527309>.
8. Dessler, G.: Human Resource Management. Pearson Education (2020).
9. Gan, C. et al.: Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening. ArXiv Prepr. (2024).
10. Hevner, A.R. et al.: Design science in information systems research. *MIS Q.* 28, 1, 75–105 (2004).
11. Hu, Q.: Unilever’s Practice on AI-based Recruitment. *Highlights Bus. Econ. Manag.* (2023). <https://doi.org/10.54097/hbem.v16i.10565>.
12. Joshi, A. et al.: Strategic Adoption of Artificial Intelligence for Human Resource Management Practices Transforming Healthcare Sector. 3, 3, 151–163 (2024). <https://doi.org/10.58818/ijems.v3i3.133>.
13. Lukaszewski, K.M.: Will the use of AI in human resources create a digital Frankenstein? *Organ. Dyn.* 53, 1, (2024). <https://doi.org/10.1016/j.orgdyn.2024.101033>.
14. Mujtaba, D.F.: Ethical Considerations in AI-Based Recruitment. 019 IEEE Int. Symp. Technol. Soc. ISTAS. 2019, (2019). <https://doi.org/10.1109/ISTAS48451.2019.8937920>.
15. Novais, L.: CurEval - Curriculum Evaluation. Instituto Superior de Engenharia do Porto (2023).
16. Page, M.J. et al.: The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ.* 372, n71 (2021). <https://doi.org/10.1136/bmj.n71>.
17. Peffers, K. et al.: A design science research methodology for information systems research. *J. Manag. Inf. Syst.* 24, 3, 45–77 (2007).
18. Pessach, D.: Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decis. Support Syst.* 134, (2020). <https://doi.org/10.1016/j.dss.2020.113290>.
19. Roxburgh, E., Hansen, K.: Bias in Recruitment and Selection. *Cent. Labour Employ. Work.* 0, 1–3 (2015).
20. Schmidt, F.L., Hunter, J.E.: The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychol. Bull.* 124, 2, 262–274 (2016).
21. Tilmes, N.: Disability, fairness, and algorithmic bias in AI recruitment. *Ethics Inf Technol.* 24, (2022). <https://doi.org/10.1007/s10676-022-09633-2>.
22. Vaswani, A. et al.: Attention Is All You Need. *Adv. Neural Inf. Process. Syst. NIPS.* 30, 5998–6008 (2017).
23. Vivek, R.: Enhancing Diversity and Reducing Bias in Recruitment Through AI: A Review of Strategies and Challenges. 2, 4, 0101–0118 (2023). <https://doi.org/10.47813/2782-5280-2023-2-4-0101-0118>.
24. Waller, M. et al.: Bias Mitigation Methods: Applicability, Legality, and Recommendations for Development. *J. Artif. INTELLIGENCE Res.* 81, 1043–1078 (2024).
25. Zhou, R.: Empirical Study and Mitigation Methods of Bias in LLM-Based Robots. *Acad. J. Sci. Technol.* (2024). <https://doi.org/10.54097/re9qp070>.