Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Toward Evidence-Based Applicant Tracking: A Framework for Evaluating AI-Assisted Recruitment Systems

Samiyullaha Sayyed

Independent Researcher

Abstract

Most evaluations of AI-assisted hiring systems still ask the same question: does this resume resemble the job posting? This paper argues that question is the wrong starting point. It proposes an evidence hierarchy—claim, then experience, then evidence, then assessment, then verified competency—where a resume statement counts for progressively more only as it is backed by progressively stronger proof, and argues that AI screening systems should be validated against both fairness and demonstrated competency jointly, not fairness alone. Companies now lean on Applicant Tracking Systems (ATS) to get through the sheer volume of resumes that online recruiting produces; that much is settled. What's less settled is whether these systems find people who can actually do the job, or whether they mostly find resumes that happen to echo the job posting's own wording. Generative AI has pulled this question into sharper focus. Job seekers can now use AI to rewrite a resume for almost any posting in seconds, which raises a fair question for employers: is a high match score still worth anything once both sides of the hiring process are running AI? Some of it is harmless polish. Grammar gets cleaner, relevant experience gets described more clearly. But the same tools can just as easily dress up a thin resume to look like a strong one. Separately, published audits of language-model-based resume screening have already found real demographic bias in these systems, which means switching from keyword search to a fancier semantic or LLM-based matcher does not, by itself, fix fairness. This paper does not report a completed experiment. It sets out a way to test these questions properly: an evaluation approach that looks past resume-to-job-description similarity and asks instead about candidate evidence, competency assessment, structured interviews, human review, fairness, and what happens after the person is actually hired. It proposes a Human-in-the-Loop Evidence-Based Candidate Evaluation Framework (HLECEF), lays out the research questions and hypotheses the framework is meant to test, and argues that ATS quality should ultimately be judged against demonstrated ability and on-the-job performance, not resume similarity.

Keywords: Applicant Tracking System, ATS, Artificial Intelligence, Generative AI, Resume Screening, Recruitment, Candidate Selection, Algorithmic Bias, Human-in-the-Loop, Skills-Based Hiring, AI Ethics, Recruitment Fraud, Semantic Matching, Large Language Models.

1. Introduction

Can an AI-assisted ATS actually tell the difference between a resume that sounds right and a candidate who is right? That question runs through everything that follows, and it exists because of a structural gap:

a candidate has real skills, real experience, real judgment built over years, while a resume is just a page or two describing some of that. Resume quality and candidate quality are not the same thing, and an ATS score is not the same thing as candidate capability either.

Hiring used to follow a fairly linear path: a company defines a role, posts it, receives applications, screens resumes, interviews a shortlist, and makes an offer. Online applications broke that flow open. The volume got too large for people to handle alone, so organizations turned to ATS platforms, ranking algorithms, and now large language models (LLMs) to help sort through everything.

The goal sounds simple enough: find the people who are actually qualified. But there is a gap between what a system can measure and what it is trying to measure. That gap has always existed, but it used to be relatively small and stable.

Generative AI is stretching that gap in both directions at once. Employers can screen applications with AI. Candidates can optimize their resumes with AI. Both sides are now running algorithms against each other, often without either side fully realizing it. A candidate who figures out what an ATS is looking for can shape their resume around it, and the resulting document might score well even if the underlying experience is thin. The reverse happens too: a genuinely strong candidate can get buried because their resume uses different words than the job posting does. One error quietly excludes people who deserve a look; the other quietly advances people who do not. That tension is the starting point for this paper.

2. Problem Statement

Take a concrete example. A company posts a role that needs Python, cloud computing, Kubernetes, Docker, CI/CD, microservices, and database administration. Candidate A has five years doing exactly this kind of work, but writes about it in plain, general language: designed and ran distributed production platforms, used container orchestration, automated the deployment pipeline. Candidate B has much less hands-on experience but submits an AI-polished resume that name-drops every term on the list: Python, AWS, Kubernetes, Docker, CI/CD, microservices, DevOps, cloud architecture.

A keyword-driven screener will likely rank Candidate B higher, even though Candidate A is the stronger hire. That is the core problem this paper is built around: a higher match score does not necessarily mean a more competent candidate. Build a recruitment system around textual similarity and you risk optimizing for the wrong thing entirely.

3. Research Motivation

Five things are pushing this problem to the front of the queue.

3.1 Application Volume Keeps Growing

Manual review simply does not scale to the number of applications many postings now receive.

3.2 Automated Screening Is Now the Default

Most applications are filtered by an ATS or an AI tool before a person ever reads them.

3.3 Generative AI Is Cheap and Fast

Candidates can now generate a tailored resume for every posting in minutes, which was not realistic even a few years ago.

3.4 Bias Concerns Are Backed by Evidence, Not Just Intuition

This is not a hypothetical worry. Wilson and Caliskan [1] ran a peer-reviewed audit of over 500 resumes and job descriptions across nine occupations and found real, statistically significant racial and gender gaps in language-model-based resume screening. Resumes with White-associated names were favored in 85.1% of the racial comparisons they tested. Resumes with female-associated names were favored in only 11.1% of the gender comparisons. Those are not small effects.

3.5 Regulators Are Paying Attention

The EU's AI Act puts recruitment and candidate-selection systems in its high-risk category [2]. The U.S. Equal Employment Opportunity Commission has published its own guidance on AI use in hiring, including how it intersects with disability discrimination law [3].

Put these five together and recruitment AI stops looking like a narrow HR-tech question and starts looking like a real technical, ethical, and regulatory problem.

4. Literature Review

4.1 Applicant Tracking Systems

ATS platforms started out as workflow tools: collecting applications, storing resumes, scheduling interviews, tracking candidates through a pipeline. Over time they picked up NLP and AI features, and it's useful to think of that evolution in four rough stages: plain document management, keyword search, semantic matching, and now AI or LLM-based evaluation. Each stage is more capable than the last, but capability and validity are different things. A more sophisticated system is not automatically a more accurate one.

5. Keyword Matching Versus Semantic Matching

Older ATS filtering leans on plain word overlap between a resume and a job posting. The simplest version of that is just the share of required terms the resume happens to contain, shown in (1).

Match = Number of Matching Terms / Required Terms (1)

Semantic matching tries to go a step further and catch ideas that are phrased differently. A posting asking for “cloud infrastructure management” might get matched to a resume describing “managing AWS and Azure production environments,” even though the two phrases don't share a single keyword.

But semantic matching brings its own problem: it can infer more than the text actually supports. If a resume says the candidate “worked with Docker,” a semantic model might read that as evidence of an “experienced container-platform engineer” — a much bigger claim than the original sentence makes. Keeping inference and evidence separate matters no matter which matching method is used.

6. Evidence From Prior Research

Wilson and Caliskan [1] built their audit around more than 500 resumes and roughly 500 job descriptions spanning nine occupations, using language-model retrieval and systematically swapping in names associated with different races and genders. They found real gender and racial gaps, effects that compounded at the intersection of race and gender, and an influence from resume length and how strongly

a name signaled a demographic group. The headline numbers bear repeating: White-associated names were favored 85.1% of the time in racial comparisons, and female-associated names only 11.1% of the time in gender comparisons.

What makes this study worth dwelling on is that it did not rely on simple keyword search at all. It tested language-model-based retrieval directly, and the bias showed up anyway. That's the important takeaway for anyone building next-generation ATS tools: swapping keyword search for embeddings or LLMs does not automatically fix bias. It can just make the bias harder to see.

7. Algorithmic Bias in Recruitment

Bias can creep into a recruitment algorithm at more than one point: in the historical hiring data used to train it, in the model itself, or in how its predictions get turned into a selection decision. If past hiring decisions were skewed, a model trained on that history will likely reproduce the skew. But bias doesn't need biased training labels to show up. Word embeddings, name-based signals, the prestige of a candidate's school, geography, gaps in employment history, resume length, writing style — all of these can act as stand-ins for protected characteristics even when nobody intended them to.

A systematic review of algorithmic discrimination in HR settings makes the same point: these systems are not neutral by default, and they need ongoing auditing [4]. Other researchers have proposed practical ways to reduce hiring bias [10], questioned the assumed neutrality of hiring algorithms [8], and run their own audits of demographic bias in AI-driven resume evaluation [11], [12]. NIST also maintains general public guidance on AI risk management that's relevant here [9].

8. Related Industry Systems: Claim Verification in Practice

A related industry system, AlteraSF [13], illustrates one concrete way claim verification can be operationalized. The platform extracts factual claims directly from resumes, generates dynamic follow-up questions to probe them, and scores responses along two quantitative axes — claim validity and job fit — supplemented by a qualitative integrity-signal check that flags linguistic patterns associated with AI-assisted or copied answers. Piloted across six job families and roughly 1,700 applications, the system reports substantial reductions in screening time. HLECEF shares the underlying premise that a resume claim needs to be tested rather than taken at face value, but the two differ in scope. AlteraSF operates as a standalone verification layer aimed primarily at compressing screening time. HLECEF instead treats verification as one stage inside a longer pipeline that keeps a human as the terminal decision-maker, weighs fairness and competency jointly rather than as separate passes, ties evaluation back to a defined competency rubric, and feeds post-hire outcomes back into the model rather than closing the loop at the hiring decision. Put differently, AlteraSF is built to answer “is this specific claim true”; HLECEF is built to answer “should this person be hired, and what evidence justifies that call.”

9. Generative AI and Resume Optimization

Generative AI changes what's possible on the candidate's side of the process. A candidate used to write one resume and reuse it. Now they can generate a fresh, tailored version for every single posting, almost for free. None of that is inherently a problem. AI can genuinely help with grammar, structure, readability, translation, formatting, and putting real experience into clearer words. Where it becomes a problem is when AI starts changing the facts on the page instead of just how those facts are presented.

10. Genuine Optimization Versus Fabrication

Two simple relationships capture the difference, shown in (2) and (3).

Truth + AI Editing = Improved Representation (2)

Limited Experience + AI Fabrication = False Qualification (3)

Fabrication here can mean invented jobs, invented projects, invented certifications, technical skills that were never actually used, inflated years of experience, achievements that never happened, or references that don't check out. This is a hard problem for ATS design specifically because a system can be very good at spotting expected words and very bad at telling whether the candidate can actually back those words up.

11. The AI Recruitment Arms Race

You can picture the interaction as a loop. The employer's job description gets run through an ATS or AI screen, which produces a ranking. Candidates respond by optimizing their resumes with AI to climb that ranking. The employer's system adapts, or gets replaced with something sharper. The candidates adapt again. Nobody planned this arms race, but the incentives point that way on both sides, and the risk is that hiring turns into a contest between two AI systems rather than an actual evaluation of who can do the work.

12. Research Gap

Plenty of research already covers algorithmic bias, automated recruitment, resume screening, language-model bias, fairness, and human-AI decision making. What gets less attention is the line between matching a resume and actually being competent for the job. Most evaluation still comes down to matching accuracy, ranking quality, retrieval performance, efficiency, and fairness — all useful, but all describing how well a system reproduces a proxy for competence rather than competence itself. This paper's contribution is putting candidate competency evidence on equal footing with those more familiar metrics.

13. Research Questions

1. How often does exact keyword matching exclude qualified candidates whose resumes use different but equivalent terminology?
2. Does semantic or LLM-based matching find more qualified candidates than keyword-based screening?
3. Does generative-AI resume optimization actually raise resume-to-job-description similarity scores?
4. Does a higher similarity score correspond to demonstrated competency, or not?
5. What are the false-positive and false-negative rates for each screening approach?
6. Does human review reduce AI-driven ranking bias, or does it sometimes make it worse?
7. What signals can organizations realistically use to catch fabricated qualifications?
8. What recruitment architecture strikes a reasonable balance between automation, efficiency, fairness, and human accountability?

14. Research Objectives

1. Evaluate keyword-based resume matching.
2. Evaluate semantic and LLM-based resume matching.
3. Measure how much AI-assisted resume optimization moves match scores.
4. Measure false-positive and false-negative rates in candidate selection.

5. Build an evidence-based candidate-evaluation framework.
6. Set out ethical and governance requirements for AI-assisted recruitment.

15. Research Hypotheses

H1. Keyword-based ATS screening has a higher false-negative rate than semantic matching when qualified candidates describe their experience using different terms than the job posting.
H2. Generative-AI resume optimization measurably increases resume-to-job-description similarity.
H3. Resume-to-job-description similarity correlates more weakly with demonstrated competency than a proper competency assessment does.
H4. Evidence-based candidate assessment produces more precise candidate selection than resume matching alone.
H5. Human-in-the-loop evaluation produces fewer selection errors than fully automated ranking.
H6. AI-generated resume optimization raises the risk of qualification misrepresentation when there's no factual verification step.
H7. AI screening systems need to be evaluated on both fairness and task competence, not fairness alone.

16. Proposed Research Methodology

This is where the paper shifts from argument to design. The plan below is a mixed-method setup: quantitative measurement of ranking and classification performance, paired with qualitative work on how recruiters actually reason through decisions, how candidates get represented, what AI explanations look like in practice, and what ethical concerns come up along the way. None of the numbers in this paper are results — they are a specification for a study that still needs to be run.

16.1 Dataset Design

A workable dataset would need somewhere between 300 and 500 job descriptions, spread across occupations such as software engineering, healthcare, finance, human resources, sales, marketing, education, project management, data science, and cybersecurity. Paired with those, 1,500 to 3,000 candidate resumes spanning a real range of quality levels. Where real resumes aren't available or aren't appropriate to use, carefully constructed synthetic resumes built around realistic competency profiles can substitute.

16.2 Resume Variants

Each candidate would ideally have several parallel versions of their resume on file: Version A — Natural resume (the candidate's own, unedited description); Version B — Keyword-aligned resume (the same real experience, rewritten using the job posting's own terms); Version C — Semantic resume (the same real experience again, but in different words that mean the same thing); Version D — AI-optimized resume (an AI-generated version built strictly from real, verified candidate information); Version E — Controlled fabrication condition (a research-only version with some unsupported claims added in, used only inside an approved ethics framework and never sent to a real employer).

16.3 Experimental Groups

Four screening approaches would be compared head to head: keyword-based ATS (Model 1), semantic embeddings (Model 2), an LLM-based evaluator (Model 3), and the HLECEF framework proposed in this paper (Model 4).

16.4 Ground Truth

The hardest methodological choice in a study like this is deciding what counts as "qualified" in the first place. A resume claiming a skill is not proof of that skill. Ground truth needs to come from a defined competency rubric that ties each required skill to a specific way of testing it, along the lines of Table 1.

Competency	Evidence Source
Python	Technical assessment
Cloud computing	Scenario-based assessment
Kubernetes	Practical task
Database administration	Technical questions
Leadership	Structured interview
Communication	Structured interview
Project management	Case study

Table 1: Illustrative Competency Rubric Mapping Required Skills to Evidentiary Methods

Under this design, ground truth means demonstrated competency, not resume keywords — and that swap is really the whole point of the framework.

16.5 Candidate Scoring Model

A composite competency score for each candidate could take the form in (4),

$$C = w1T + w2P + w3I + w4E \quad (4)$$

where T is the technical-assessment score, P the practical or project-assessment score, I the structured-interview score, and E the verified-experience score. The weights w1 through w4 would need to be tuned per job category rather than fixed globally.

17. Evaluation Metrics

17.1 ATS Performance Metrics

Standard classification metrics apply here, shown in (5) through (7), where TP, FP, and FN are qualified candidates correctly selected, unqualified candidates incorrectly selected, and qualified candidates incorrectly rejected.

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (7)$$

17.2 Fairness Metrics

Where demographic data is available and ethically appropriate to use for research, fairness analysis can include a group-level selection rate and how that rate compares to a reference group, shown in (8) and (9).

$$SRg = Selectedg / Applicantsg \quad (8)$$

$$SRR = SR(group) / SR(reference) \quad (9)$$

Demographic parity, equal opportunity, gaps in false-negative rates between groups, and intersectional effects all belong in this analysis too. But fairness alone is not enough of a test. A system can look

statistically balanced and still be bad at finding competent people. That's why this paper argues for checking both fairness and competency together, covered further in Section 22.

18. Proposed HLECEF Framework

HLECEF treats the hiring pipeline as a series of evidence-gathering steps rather than one big ranking pass. It starts with the job description, moves through requirement mapping and semantic ATS screening to form a candidate pool, then splits into two parallel checks on that pool: claim analysis and anomaly analysis. Those two feed into a skills assessment, then a structured interview, then human review, then a final decision, and finally into tracked post-hire outcomes that loop back to improve the model over time. The one rule that matters most: no single automated step gets to make the final call on its own. Every step contributes evidence; a person makes the decision.

19. Why Human-in-the-Loop Oversight Is Not Automatically Sufficient

Adding a human reviewer is often treated as the fix for AI bias, but people are not immune to being steered by an AI recommendation. A recent study with more than 500 participants found that human resume-screening decisions shifted noticeably when people worked alongside an AI system that showed biased preferences [5]. So pairing AI with a human reviewer doesn't automatically produce a fair result. The human step itself needs real design: blind initial review, structured scoring, independent assessment before anyone sees an AI recommendation, a requirement to explain overrides, and regular audits.

20. AI as an Evidence Assistant, Not the Final Judge

In this framework, AI has a fairly narrow job. It can parse applications, summarize content, spot relevant and transferable skills, compare a candidate's material against the stated requirements, flag inconsistencies, suggest good assessment questions, and point out where evidence is missing. What it should not do is decide, on its own, that a given person deserves the job. That call has to combine evidence from several sources, with a human making the final decision and owning it.

21. Qualification Verification and the Evidence Hierarchy

There's a real difference between a claim (“Kubernetes expert”), a supported claim (“managed a Kubernetes production cluster with 40 nodes”), and a verified claim (a technical assessment that actually shows Kubernetes competency). Ordering these gives an evidence hierarchy — claim, then experience, then evidence, then assessment, then verified competency — where each step up carries more weight than the last. That hierarchy is one of the more useful pieces of this framework.

22. Detecting AI-Assisted Fabrication

It's a mistake to treat AI-assisted writing itself as suspicious — too many people use these tools for that to mean anything on its own. The better approach is watching for claims that don't hold up: skill combinations that are unusually broad, dates that don't line up, seniority jumps that don't make sense, certifications nobody can verify, project scale that seems implausible, employment history that contradicts itself, technical claims that don't match the stated responsibilities, or a candidate who can't explain a project they claim to have led. The real question isn't whether AI touched the resume. It's whether the candidate can back up what it says.

23. Application Automation and Volume Inflation

AI agents can now handle most of the application pipeline on their own — finding postings, classifying them, customizing a resume, writing a cover letter, filling out the form, and hitting submit — at a scale no human could match. That opens the door to application-volume inflation: more applications coming in without a matching increase in genuinely qualified applicants. That just pushes more work onto whatever screening system sits downstream, and it's another reason to build around evidence rather than volume.

24. Ethical Framework

A responsible recruitment AI system needs to hold up against at least eight principles: Transparency (candidates and recruiters should know where AI is being used); Accountability (the organization stays responsible for its hiring decisions, automation or not); Fairness (systems get tested for discriminatory outcomes on a regular basis); Explainability (organizations can actually explain what drove a selection decision); Privacy (candidate data is protected properly throughout the process); Security (documents candidates submit are treated as untrusted input); Human oversight (high-stakes decisions don't rest on automated ranking alone); Competency validation (the claims that matter get checked against evidence before anyone relies on them).

These line up reasonably well with NIST's AI Risk Management Framework, which covers validity and reliability, safety, security, accountability, transparency, explainability, privacy, and managing harmful bias [6], and with NIST's Generative AI Profile, which addresses risks specific to generative systems [7].

25. Regulatory Context

25.1 European Union

The EU AI Act (Regulation (EU) 2024/1689) puts recruitment and selection systems — including anything that filters applications or evaluates candidates — in its high-risk category [2]. That's a meaningful signal: recruitment AI isn't treated as an ordinary productivity tool, because it can genuinely affect someone's access to a job, and the obligations reflect that.

25.2 United States

The U.S. EEOC has its own guidance on AI in employment selection, including how it connects to disability discrimination [3]. Any organization deploying recruitment AI should be checking that system against actual employment-discrimination law, not assuming it sits outside legal scrutiny just because it's automated.

26. Proposed Statistical Analysis

Depending on how the dataset turns out, a few analyses would be worth running: a Chi-square test for differences in selection rates across groups; logistic regression to estimate selection probability from candidate and application features; ANOVA to compare average scores across the four screening methods; correlation analysis across ATS score, resume similarity, assessment score, interview score, and post-hire performance; and regression analysis to test whether ATS score actually predicts job performance.

That last test matters more than the others. If ATS score barely predicts post-hire performance, that's a strong signal that organizations are leaning too heavily on ranking as a proxy for quality. If it predicts performance well, that's evidence the screening system is actually doing its job.

27. Proposed Reporting Tables

Tables 2 and 3 show the shape the results should eventually take. Every cell here is a placeholder. This paper has not run the experiment, so none of these numbers exist yet — they should only get filled in once real data has been collected and analyzed.

Approach	Precision	Recall	F1	FPR	FNR
Keyword ATS	TBM	TBM	TBM	TBM	TBM
Semantic Matching	TBM	TBM	TBM	TBM	TBM
LLM Evaluation	TBM	TBM	TBM	TBM	TBM
HLECEF	TBM	TBM	TBM	TBM	TBM

Table 2: Proposed Screening-Performance Comparison (TBM = To Be Measured)

Model	Gender Diff.	Race Diff.	Intersect. Diff.	Signif.
Keyword ATS	Exp.	Exp.	Exp.	Exp.
Semantic Model	Exp.	Exp.	Exp.	Exp.
LLM	Exp.	Exp.	Exp.	Exp.
HLECEF	Exp.	Exp.	Exp.	Exp.

Table 3: Proposed Fairness-Analysis Comparison (Exp. = Experimental / To Be Determined)

This kind of side-by-side fairness comparison matters because of what Wilson and Caliskan [1] already found: real demographic gaps in LLM-based resume retrieval, and intersectional effects that a single-axis analysis would have missed entirely.

28. Discussion

Automation isn't really the problem here. Optimizing for the wrong target is. If a screening system's real criterion is “find resumes with these words in them,” candidates will write resumes with those words in them — that's just a rational response to the incentive. If the criterion is “find candidates who can demonstrate these skills,” the system has to look at actual evidence instead of surface text. Optimizing a proxy is not the same as achieving the goal the proxy was supposed to represent. Resume similarity is a proxy. Candidate capability is the actual goal.

29. Implications

29.1 For HR Professionals

Recruiting teams would do well to shift from resume-first thinking to evidence-first thinking. Instead of asking whether a resume mentions a given term, ask what evidence shows the candidate can actually do that part of the job.

29.2 For AI Developers

Vendors selling recruitment AI should be willing to share real information about their systems: model version, how it was evaluated, known limitations, fairness-testing results, error rates, how human review fits in, audit capability, data-governance, and security. This is a high-impact system, and it should be treated like one.

29.3 For Candidates

Using AI to improve grammar, restructure real experience, sharpen how achievements are described, or find the right terminology is fine. Using it to invent qualifications is not. AI can improve how something is communicated. It shouldn't be used to manufacture the underlying competence.

29.4 For Organizations

It's worth being careful about the leap from "the ATS picked these 20 people" to "these are the 20 best candidates." The more accurate statement is narrower: the ATS ranked these 20 candidates highest under its current model and criteria. That's a claim about the system's output, not an independent claim about who's actually the best fit.

30. Limitations

This paper has real limits worth being upfront about. ATS platforms are proprietary and vary a lot between vendors, so no single evaluation generalizes to all of them. Public evidence can't cover every system on the market. Building the dataset and competency ground truth this framework calls for would take real time and money. Post-hire performance is genuinely hard to measure consistently across different jobs. Fairness research involving demographic attributes needs careful, ethically grounded handling. Generative AI is also changing fast enough that candidate and employer behavior could shift meaningfully over the course of a multi-stage study. And even a well-designed controlled experiment won't perfectly reproduce how messy real organizational hiring actually is.

31. Future Research Directions

- Agentic AI-driven application systems.
- Automated qualification verification.
- AI-generated credential fraud.
- Resume prompt injection aimed at AI-based screeners.
- Multimodal candidate assessment.
- AI-conducted interviews.
- Cross-cultural recruitment bias.
- Accessibility and disability-related impacts of AI screening.
- Longitudinal post-hire performance analysis.
- Explainable candidate ranking.
- AI-to-AI recruitment interactions.
- Privacy-preserving candidate evaluation.

One question worth chasing in particular: can a recruitment AI system tell whether a candidate actually has a claimed skill without relying entirely on how the candidate describes it themselves?

32. Conclusion

A resume is evidence about a candidate, not the candidate. That single fact is what this paper has tried to keep in view throughout. ATS platforms genuinely help organizations deal with large application volumes, and AI can meaningfully improve document processing, semantic matching, candidate discovery, and recruiter productivity — none of that is in question. What generative AI changes is the size and shape of the gap between resume and reality: candidates keep optimizing their resumes for ATS systems while

employers respond with sharper screening AI, an arms race where the risk is that selection ends up optimized for beating the screen rather than for finding people who can do the work.

Wilson and Caliskan's audit [1] shows real demographic disparities in language-model-based resume screening, which is a clear signal that moving from keywords to sophisticated language models doesn't fix bias on its own. Regulatory moves in the EU and US [2], [3] reflect the same conclusion from a different direction: AI-assisted recruitment is increasingly treated as a high-impact use case that needs governance, risk management, transparency, and human oversight.

This paper argues that recruitment screening needs to move through keyword matching, semantic matching, and AI evaluation, and then keep going — toward evidence verification, competency assessment, and a human decision at the end of the chain. HLECEF is one attempt at putting evidence and competency at the center of that process, anchored by an evidence hierarchy that treats a bare claim and a verified competency as fundamentally different kinds of proof. The real test of an ATS isn't how well it can match a resume to a job description. It's how well it can find the people who can actually do the job.

References

1. Wilson, K., & Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 7(1), 1578–1590.
2. European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Annex III, Employment and Worker-Management High-Risk AI Systems.
3. U.S. Equal Employment Opportunity Commission. Artificial intelligence and automated systems in employment selection. EEOC Guidance and Resources.
4. Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13, 795–848.
5. Wilson, K., Sim, M., Gueorguieva, A.-M., & Caliskan, A. (2025). No thoughts just AI: Biased LLM hiring recommendations alter human decision making and limit human autonomy. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 8(3), 2692–2704.
6. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. National Institute of Standards and Technology.
7. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. National Institute of Standards and Technology.
8. Mann, G., & O'Neil, C. (2016). Hiring algorithms are not neutral. Harvard Business Review.
9. National Institute of Standards and Technology. AI Risk Management Framework and AI Resource Center.
10. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*).
11. Webster, K. T. (2025). Fairness is not enough: Auditing competence and intersectional bias in AI-powered resume screening. arXiv preprint.

12. Wen, A., Patil, T., Saxena, A., Fu, Y., O'Brien, S., & Zhu, K. (2025). FAIRE: Assessing racial and gender bias in AI-driven resume evaluations. arXiv preprint arXiv:2504.01420.
13. Lee, E., Worley, N., & Takatsuji, K. (2025). Quantifying truth and authenticity in AI-assisted candidate evaluation: A multi-domain pilot analysis. arXiv preprint arXiv:2511.00774.