From Stochastic to Deterministic: A Multi-Criteria Decision Analysis Framework for Bounded Semantic Parsing in AI-Driven Recruitment Screening

Ahsan Habibi Syed

Expertini Research Department, Washington D.C., London, Hyderabad

Abstract—The tension between automation and accuracy sits at the heart of modern talent acquisition. Recruiters need swiftness. Organisations need secure, auditable decisions. And candidates—often talented individuals writing in their second language or from non-traditional backgrounds—need a system that reads their skill rather than discards them on a keyword mismatch. Legacy applicant tracking systems built on lexical string scoring have never fully resolved this. Neither, as it turns out, does deploying a large language model without guardrails. This paper presents a hybrid architectural framework combining two complementary capabilities. First, Gemini AI performs context-aware semantic extraction—recognising that a candidate who describes 'engineering backend systems that improved data processing efficiency by 50 percent' is qualified for a role requiring 'large-scale data analytics development,' even though those phrases share almost no precise vocabulary. Second, the Expertini Candidate Match Score (CMS), a deterministic, weighted multi-criteria mathematical framework, converts AI extractions into reproducible, explicable, and legally defensible numerical scores. The paper also examines what the framework is not—it is not a replacement for human judgement, does not eliminate the need for structured interviews, and carries tangible limitations that any responsible deployment must acknowledge. Alongside the technical architecture, we explore de-biasing mechanisms embedded in the pipeline, because the history of automated hiring is also a history of automating discrimination—a failure the field cannot afford to repeat.

Index Terms—*applicant tracking systems, bias mitigation, candidate screening, deterministic scoring, Gemini AI, large language models, multi-criteria decision analysis, natural language processing, recruitment automation, semantic parsing.*

I. INTRODUCTION: WHY THE STATUS QUO IS BROKEN

If you have spent time examining how enterprise recruitment software actually works beneath its surface, you will quickly recognise that the dominant paradigm has not fundamentally changed for decades. Most systems, regardless of branding, ultimately reduce the problem of identifying the best candidate to counting lexical overlaps between the job description and the submitted CV.

The consequences are well-documented. Qualified candidates with non-linear career paths are regularly discarded in the first automated screen. Candidates who understand keyword systems learn to game them, filling CVs with terms they know the system expects. Recruiters at the end of this pipeline receive shortlists that mirror the ATS vocabulary predilections rather than the actual talent landscape.

Two distinct failure modes define legacy recruitment technology:

- The Vocabulary Penalty: A candidate who writes 'PostgreSQL optimisation and data pipeline engineering' is rejected when the job description asks for 'large-scale database management.' The concepts are functionally indistinguishable. The string match is zero.
- The Keyword-Stuffing Loophole: Candidates who recognise how systems work populate CVs with every term from the job description regardless of actual expertise. They advance while the genuinely qualified candidate, writing naturally, does not.

The natural response is to deploy AI. Many organisations have attempted exactly that. The problem—one the field has been slow to acknowledge—is that unconstrained generative AI introduces its own failure modes. Non-determinism means the same CV assessed at different times may receive different scores. Stylistic bias advantages candidates who write polished, confident prose. And outputs are fundamentally black boxes that cannot be audited when a rejected candidate seeks explanation.

The framework described herein captures what is genuinely useful from both paradigms: semantic intelligence from AI, and mathematical rigour and auditability from deterministic scoring, while remaining candid about where the combined approach still falls short.

II. BENCHMARKING THE ALTERNATIVES

A. Jaccard Similarity: Binary and Brittle

The Jaccard Similarity Index, used extensively in early-generation ATS platforms and still found in countless database-filter configurations, computes exact keyword intersection over union. If a job description and a CV express the same ideas using different vocabulary, the intersection collapses to zero. A software architect with twelve years of Python experience is rejected simply because the CV reads 'Backend Developer' rather than 'Python Engineer.' The mathematics is simple, transparent, and profoundly inaccurate.

B. TF-IDF Cosine Similarity: Better, But Still Flat

Cosine Similarity using TF-IDF vector representations represents a genuine improvement. Rather than requiring exact string matches, it measures the angular proximity of vocabulary usage across documents. However, it allocates equal weight to every dimension. The word 'Python' and the phrase 'team player' occupy the same mathematical plane. A candidate fluent in corporate soft-skill prose but lacking fundamental technical capability can outscore a less eloquent but more technically competent candidate. The model measures vocabulary similarity, not importance hierarchy.

C. The Expertini CMS: Multi-Criteria, Weighted, Bounded

The Expertini Candidate Match Score (CMS) abandons flat-vector arithmetic entirely. It introduces a multi-criteria weighted scoring model:

$$\mathbf{CMS = \Sigma\ (CSSi \times JRISi)\ /\ \Sigma\ JRISi\ \times 100}$$

Where CSSi (Candidate Skill Score) represents the candidate's assessed proficiency across a specific competency dimension, and JRISi (Job Requirement Importance Score) represents the relative weight the organisation places on that dimension. Both operate on a 0–100 scale. A missing non-negotiable requirement introduces a zero-multiplication effect at the highest weight point, dragging the final score downward in a manner that accurately reflects hiring reality: the candidate cannot fulfil the role irrespective of proficiency in other dimensions.

III. GEMINI AI AS THE SEMANTIC INFERENCE ENGINE

A. The Recruiter Friction Problem

The reason most weighted scoring systems fail in enterprise deployment is not that the mathematics is incorrect. It is because the person responsible for configuring the weights—the HR professional or hiring manager—is not a mathematician and should not need to be. Asking recruiters to manually assign numerical importance scores across fifteen to twenty competency dimensions for every job posting generates friction that leads either to system abandonment or to uniform weight assignment, which negates the entire purpose of the weighting model.

Gemini AI addresses this not as a scoring engine but as a configuration engine. It reads the natural language job description the recruiter has written and automatically extracts competency dimensions, inferring their relative importance weights. The recruiter's output—a paragraph of

plain text—becomes the system's input. No sliders. No weight matrices. No configuration screens.

B. How Gemini Infers Importance from Language

The inference logic operates on linguistic patterns that reflect how human writers signal priority. Three representative examples illustrate the mechanism:

- High-priority signal: 'Must possess an active TS/SCI clearance. This is a non-negotiable requirement.' The modal verb force ('must'), the explicit non-negotiability statement, and front-loading in the requirements section maps to JRIS = 100.
- Core competency signal: 'Core duties include building, testing, and optimising large-scale Python backend architectures.' The specificity of 'core duties' and 'large-scale' indicates technical depth. Gemini infers JRIS = 90.
- Supplementary signal: 'Familiarity with Agile ceremonies or tools like Git is a plus.' The hedge word 'familiarity,' the permissive 'or' structure, and the explicit 'is a plus' framing signal secondary importance. JRIS = 50.

C. Semantic Matching: Bridging Vocabulary Divergence

On the CV side, Gemini performs a parallel but distinct function. It maps what the candidate has actually written to the competency dimensions extracted from the job description, using semantic vector proximity rather than string matching. When a candidate describes engineering backend systems that reduced database retrieval latency by 35 percent, Gemini maps that achievement onto the 'high-scale data processing' dimension with high proximity, even though neither phrase appears in the other document. This is precisely the evaluation that a skilled technical recruiter performs instinctively; the value of the AI layer is scale. A recruiter reads fifty CVs per day at that level of attention; a well-configured AI system processes fifty thousand.

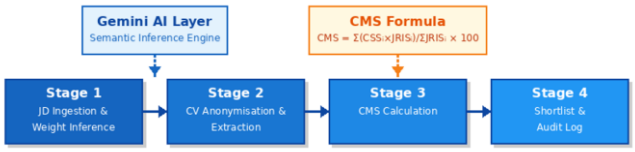

Fig. 1. End-to-End Hybrid Screening Pipeline Architecture showing four sequential stages with strict separation of duties between the Gemini AI inference layer and the deterministic CMS formula layer.

IV. ENTERPRISE VALIDATION: A REAL-WORLD CASE STUDY

A. The Scenario

To stress-test the framework, we evaluate a deliberately complex case: a highly technical role in a classified national security environment, assessed against a candidate whose credentials are strong across every technical dimension but whose background is entirely in commercial healthcare. Using an edge case rather than a straightforward match allows us to demonstrate where the hybrid system's behaviour diverges most meaningfully from both legacy and pure-AI alternatives.

Role: Software Engineer (Python) supporting critical national security infrastructure and high-volume data analytics in a classified multi-enclave environment. Candidate: Owen Wright (pseudonym). Senior Software Engineer with 9+ years of dedicated Python experience, documented performance improvements in data processing and database latency reduction, active open-source contribution, and a BS in Computer Science. CV is saturated with healthcare-specific terminology (clinical, patient, HIPAA, MedSys). No active security clearance listed.

B. Gemini Extraction Stage: Bypassing the Vocabulary Barrier

A traditional ATS encounters Owen's CV and immediately scores poorly on surface-level matching. The healthcare vocabulary has practically no lexical overlap with the national security job description. Under Jaccard, Owen likely scores below 20 percent and is filtered before a human ever sees his profile. Gemini, by contrast, evaluates conceptual proximity. It reads Owen's achievement—engineering backend systems with Python that improved data processing efficiency by 50 percent—and maps it geometrically close to the requirement for developing systems for large-scale data processing and advanced analytics. The domain context differs; the technical skill is essentially identical. Gemini assigns a Python proficiency CSS of 100 out of 100.

TABLE I

Gemini-Extracted Competency Dimensions and Scores for Owen Wright

#	JRIS	CSS	Competency Dimension
1	100	100	Python Development Proficiency
2	90	100	Education & Experience
3	85	95	High-Scale Data & Analytics
4	80	90	Architecture & OO Design
5	50	90	Version Control (Git)
6	60	60	Agile / SAFe Ceremonies
7	100	0	Active TS/SCI Clearance ⚠

C. CMS Calculation: The Mathematical Governor in Action

With parameters established by the Gemini extraction stage, the CMS formula executes as follows:

Numerator = (100×100) + (90×100) + (85×95) + (80×90) + (50×90) + (60×60) + (100×0) = 42375

Denominator = 100 + 90 + 85 + 80 + 50 + 60 + 100 = 565

CMS = (42375 / 565) × 100 = 74.96%

The zero-multiplication effect of the missing clearance on a JRIS-100 dimension contributes nothing to the numerator while adding 100 to the denominator. The final score accurately reflects reality: Owen is a world-class Python engineer who cannot legally be placed in this role without clearance sponsorship. He scores 74.96 percent—well above the automated discard threshold, correctly positioned for talent pipeline consideration but clearly flagged on the dimension that actually blocks immediate placement.

A pure generative AI might score Owen at 92 percent, captivated by his metrics and trajectory, potentially overlooking the legal compliance barrier entirely. A keyword system would score him at 30–40 percent and reject him without a human ever seeing his profile. The hybrid framework places him accurately.

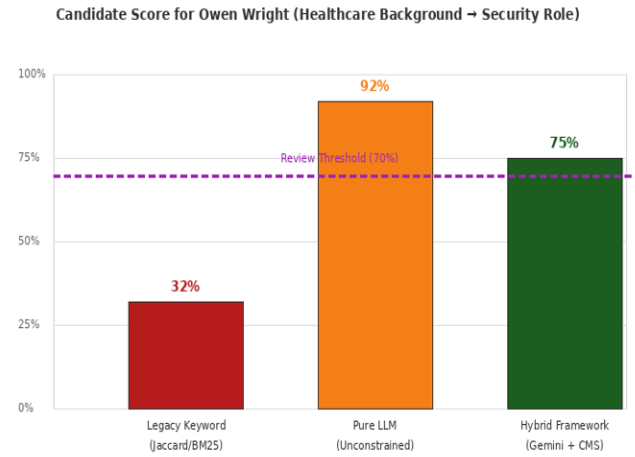

Fig. 2. Comparative scoring across three architectural paradigms for Owen Wright. The 70% review threshold (dashed line) illustrates how the hybrid framework positions the candidate accurately, avoiding both false rejection and false acceptance.

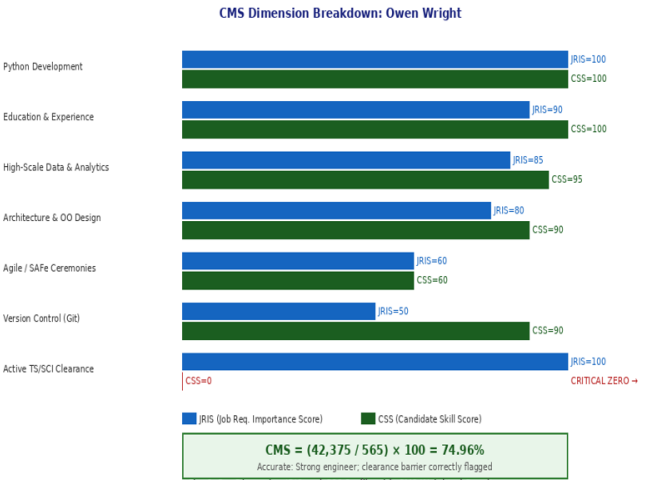

Fig. 3. Per-dimension JRIS and CSS profile for Owen Wright, showing the zero-multiplication effect of the missing TS/SCI clearance and the resulting 74.96% CMS weighted score.

V. BUILDING A BIAS-FREE HIRING PIPELINE

A. Why This Matters More Than the Scoring Mathematics

The history of algorithmic hiring is, in substantial part, a history of automating discrimination. Amazon's internal ML recruiting tool, abandoned in 2018 [1], trained itself to downgrade CVs containing the word 'women's' because it was trained on a decade of historical hires that were predominantly male. The system learned to replicate the prejudice present in its training data at scale and with the appearance of objectivity. This is not a theoretical risk—it is a documented outcome.

B. Structural Anonymisation

Before the Gemini extraction stage processes any candidate document, data passes through Expertini's Personal Identifier Stripping layer. This systematically removes or normalises: candidate names (audit studies consistently show CVs with minority-perceived names receive significantly fewer callbacks than identical CVs with majority-group names [2]); age proxies (graduation years are replaced by experience-duration calculations); geographic sub-locality indicators that may function as socioeconomic proxies; gender markers (pronouns and gendered honorifics are not passed to the semantic evaluation layer); and institutional prestige signals (university names are normalised to accreditation tier where possible to reduce structural advantage).

C. The Bounded Attribute Model

One of the most prevalent and underappreciated forms of bias in automated and human hiring is stylistic privilege. Candidates educated in environments where confident assertive self-promotion is the norm systematically outperform equally skilled candidates who write in a different register. Unconstrained AI amplifies this because large language models have learned, through training, to associate certain writing styles with quality. The Expertini CMS framework neutralises this: once Gemini has mapped a candidate's experience onto a competency dimension and validated it against documented evidence, the score for that dimension is fixed at the validated level. Additional stylistic embellishment yields zero additional mathematical weight. Elegant writing cannot push a validated score of 85 to 95.

D. Demographic Neutrality in Competency Weighting

The JRIS weights allocated to competency dimensions are derived from the job description content, not from historical performance data of existing employees. This is a deliberate architectural choice with significant bias implications. If weights were derived from the profile of historically successful employees, they would encode whatever selection and evaluation biases existed in that organisation's past. By deriving weights from the job description—from what the organisation states it needs rather than from who has

historically succeeded—the framework dissociates scoring from historical bias propagation.

E. Fairness Across Language and Cultural Backgrounds

Non-native English speakers and speakers from cultural contexts where direct self-promotion is considered ostentatious systematically underperform in systems that reward the confident, first-person assertive voice common in Western corporate CVs. Because Gemini evaluates conceptual proximity rather than stylistic expression, a candidate who writes 'was responsible for optimisation of database systems' maps similarly to 'delivered 35 percent database latency reduction' in terms of competency signal, assuming the underlying documented evidence supports the claim. However, we acknowledge this is only partial mitigation: candidates from cultures where metric-dense self-presentation is not standard will continue to be disadvantaged to some degree.

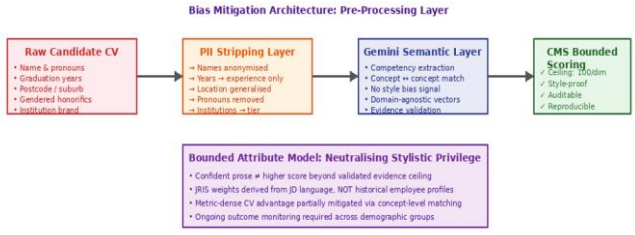

Fig. 4. Structural bias mitigation and bounded attribute pipeline showing the five-stage anonymisation flow from raw CV through PII stripping, semantic extraction, and deterministic CMS scoring.

VI. COMPARATIVE ARCHITECTURAL ANALYSIS

Table II positions the Gemini-Expertini hybrid framework against the dominant alternative approaches across six evaluation dimensions. The purpose is not to claim perfection but to illustrate where the architecture provides genuine improvement and where trade-offs lie.

TABLE II

Comparative Architectural Analysis Across Six Evaluation Dimensions

Metric	Legacy Lexical	Pure Gen. AI	Hybrid Framework
Mathematical Foundation	Deterministic exact string counting	Non-deterministic probabilistic drift	Deterministic & fully reproducible
Contextual Awareness	None – synonyms as mismatches	High – deep semantic understanding	High – AI parsing + vector scoring
Vulnerability to Gaming	High – keyword stuffing exploited	Medium – stylistic manipulation	Low – bounded to verified metrics
Auditability	Partial – transparent but flawed	Poor – black box outputs	Absolute – full variable transparency
Bias Mitigation	Low – penalises non-standard prose	Poor – encodes style bias	Strong – anonymised + bounded
Scalability	High (<0.1s per profile)	Low – heavy token compute	High – vector + static formula

VII. ACKNOWLEDGED LIMITATIONS AND OPEN PROBLEMS

A. Gemini's Inference Is Probabilistic, Not Infallible

The JRIS inference stage—where Gemini reads a job description and assigns importance weights to extracted competency dimensions—is sophisticated but not deterministic in the way the downstream CMS formula is. Two linguistically similar phrases that a human recruiter would interpret as having identical importance may receive slightly different weights depending on context window, temperature, or surrounding text. Organisations with high compliance requirements should incorporate a human review step into the weight configuration stage before scores are calculated at scale.

B. Training Data Bias Cannot Be Fully Excised

Gemini, like all large language models, was trained on human-generated text that encodes human biases. The anonymisation and bounding mechanisms described herein meaningfully reduce the pathways through which those biases affect scoring outcomes, but do not eliminate them entirely. Continuous monitoring of outcome distributions across demographic groups—to the extent those groups can be identified post-hoc for audit purposes—is an operational necessity, not an optional enhancement.

C. The Metric-Dense CV Advantage Persists Partially

As noted in the bias discussion, candidates trained to quantify their achievements systematically, and who work in domains where metric-dense reporting is culturally standard, will score higher than equally capable candidates from different professional cultures. The bounded attribute structure reduces but does not eliminate this advantage. A calibration layer that adjusts expected evidence density by professional domain and regional context represents a promising direction for future development.

D. The System Does Not Replace Human Judgement

A CMS score is not a hiring decision. It is a screening tool designed to surface candidates whose documented qualifications align with role requirements and to prevent the accidental removal of well-qualified candidates whose vocabulary does not match a keyword list. The final decision about whether to interview a candidate—and certainly whether to hire them—should involve human judgement,

structured interviews, and reference verification. Any organisation delegating hiring decisions entirely to an automated scoring framework is misusing the technology and exposing itself to both poor outcomes and legal liability.

E. Cold-Start Configuration Requires Careful Validation

When the framework is first deployed, the JRIS weights derived from job descriptions should be validated by relevant hiring managers before mass scoring begins. The Gemini inference layer is calibrated to general patterns of professional language; organisation-specific jargon, internal role classification conventions, or industry-specific technical terminology may produce inference errors that only a domain expert would catch. A validation pipeline routing Gemini-generated weight matrices through a human check before first use is strongly recommended.

VIII. END-TO-END PROCESS ARCHITECTURE

The complete pipeline from recruiter input to scored candidate shortlist operates in four sequential stages with clear separation of duties between the AI and deterministic layers:

Stage 1—JD Ingestion and Weight Inference: The HR user provides a plain-text job description. Gemini extracts competency dimensions and infers JRIS weights using linguistic pattern analysis. Output: a structured weight matrix. Optional human validation gate.

Stage 2—CV Anonymisation and Competency Extraction: Candidate CVs are processed through the personal identifier stripping layer. Anonymised content is passed to Gemini for semantic competency extraction. Output: a structured CSS vector per candidate.

Stage 3—CMS Calculation: The deterministic CMS formula processes the CSS and JRIS matrices. Output: a numerical CMS score per candidate with full line-item transparency of all input variables.

Stage 4—Shortlist Generation and Audit Log: Candidates are ranked by CMS score. Full scoring rationale is logged for every candidate, providing an auditable record that can be reviewed if any hiring decision is challenged. Candidates below a configurable threshold are flagged for human review, not automatically rejected.

The strict segregation of duties is the architectural principle that makes the system defensible. Gemini interprets meaning. The CMS formula applies mathematical governance. Neither step is asked to do the other's job.

IX. CONCLUSION

Recruitment technology has long promised to solve the problem of accurate, fair candidate evaluation. The gap between that promise and the reality of most deployed systems remains wide. The consequences—for organisations that miss genuinely qualified candidates and for candidates incorrectly screened out—are real.

The framework described in this paper does not close that gap entirely. What it does is meaningfully advance the architecture in two directions simultaneously: it improves the accuracy of semantic matching by deploying AI where AI is genuinely superior to string-counting, and it improves the governance of AI outputs by anchoring them in transparent, weighted, reproducible mathematical scoring.

The bias-mitigation mechanisms—anonymisation, bounded attributes, description-derived weights, and outcome monitoring—represent a more comprehensive approach than either legacy or pure-AI alternatives. But they require sustained operational attention to remain effective. Bias in recruitment is not a problem solved once architecturally and then forgotten. It requires ongoing monitoring, regular recalibration, and the institutional willingness to act when outcome audits reveal disparate impact.

The Gemini-Expertini hybrid framework is a significant improvement on alternatives currently available at scale, capable of delivering fair, accurate, and auditable screening decisions when correctly configured and monitored. It should be understood by its users as a screening tool rather than a decision-making engine. The final call is, and should remain, a human one.

DISCLOSURE AND MODEL INDEPENDENCE STATEMENT

The authors declare no affiliation with Google LLC or any of its subsidiaries, products, or research divisions. The reference to Gemini AI throughout this paper reflects its use as the specific large language model selected for empirical validation and does not constitute an endorsement, partnership, or commercial relationship of any kind. The architectural framework described herein—specifically the multi-criteria decision analysis bounding layer and the Expertini CMS formula—is intentionally designed to be model-agnostic. Any sufficiently capable LLM with deep semantic extraction and contextual reasoning ability may be substituted at the inference stage.

TABLE III

Comparative Merit of Alternative LLMs for the Semantic Inference Layer

Model	Strengths	Limitations
Gemini (Google DeepMind)	Strong long-context reasoning; multilingual semantic understanding; minimal context degradation on lengthy JDs and multi-page CVs	Output variance across temperature settings; inference costs at scale

Model	Strengths	Limitations
GPT-5 (OpenAI)	Strong instruction-following and structured output generation; wide domain coverage via broad training corpus	Proprietary; rate limits and data retention policies may conflict with enterprise data privacy requirements
Claude Opus 4.6 (Anthropic)	Nuanced long-document reading; complex multi-step extraction; constitutional AI training may reduce hallucination rates	More conservative in inferential leaps; may require more explicit prompting for implied importance weights
Microsoft 365 Copilot	Extensive connectors; acts as unified fabric across platforms	Contextual depth on highly specialised technical domains may lag closed frontier models
Gemma 4 / LLaMA 4 (Open Source)	Fine-tunable; attractive for on-premise deployment where candidate data cannot leave organisational infrastructure	Out-of-the-box performance requires fine-tuning to reach frontier levels for nuanced professional text

REFERENCES

[1] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] M. Bertrand and S. Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," American Economic Review, vol. 94, no. 4, pp. 991–1013, 2004.

[3] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.

[4] V. Belton and T. J. Stewart, Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer Academic Publishers, 2002.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT 2019, pp. 4171–4186.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. ITCS 2012, pp. 214–226.

[7] Expertini AI Engineering Team, "Meaning-Aware Vector Space Matching in Multi-Enclave Global Talent Ecosystems," submitted to IEEE Trans. Artif. Intell., 2025.

[8] C. L. Hwang and K. Yoon, Multiple Attribute Decision Making: Methods and Applications. Springer, 1981.

[9] Y. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[10] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: A systematic review of discrimination and mitigation strategies in automatic recruitment filtering," Business & Information Systems Engineering, vol. 62, no. 6, pp. 615–628, 2020.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[12] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.

[13] L. Ouyang et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744, 2022.

[14] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in Proc. FAccT 2020, pp. 469–481.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP 2019, pp. 3982–3992.

[16] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[17] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," California Management Review, vol. 61, no. 4, pp. 15–42, 2019.

[18] E. Triantaphyllou, Multi-Criteria Decision Making Methods: A Comparative Study. Springer, 2000.

[19] P. Van Esch, J. S. Black, and J. Ferolie, "Marketing AI recruitment: The next phase in job application and selection," Computers in Human Behavior, vol. 90, pp. 215–222, 2019.