Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence in Career Guidance and Talent Matching: A Structured Methodological Review of Models, Labels and Trustworthy Evaluation

Ayman Negm

ayman.mostafa1@msa.edu.eg

October University of Modern Sciences and Arts

Haneen Sherif

October University of Modern Sciences and Arts

Mona Farouk

Cairo University

Systematic Review

Keywords: artificial intelligence, machine learning, recommender systems, knowledge graphs, career guidance, algorithmic fairness, trustworthy evaluation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11650619/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Artificial Intelligence in Career Guidance and Talent Matching: A Structured Methodological Review of Models, Labels and Trustworthy Evaluation

Ayman Negm[1]* · Haneen Sherif[1] · Mona Farouk[1,2]

[1] Computer Systems Department, Faculty of Engineering, October University for Modern Sciences and Arts (MSA), Giza, Egypt

[2] Computer Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt

*Correspondence: Ayman Negm, ayman.mostafa1@msa.edu.eg

Abstract

Artificial intelligence (AI) is increasingly used in career guidance, occupational recommendation, and candidate-job matching, yet these applications are often treated as interchangeable despite serving different stakeholders and creating distinct downstream risks. This structured methodological review examines how AI methods are evaluated across career matching and algorithmic trustworthiness. Two semantic searches identified 980 unique records, of which 181 studies published in 2020-2026 met the extraction criteria under screening. We purposively selected 19 primary studies for in-depth analysis. Although the field has grown rapidly (62.4% of candidate studies appeared in 2025-2026), the close-read studies show that model performance depends heavily on proxy labels and validation design. High offline performance (area under the curve (AUC) > 0.92) did not consistently translate into robust outcomes: performance degraded for job seekers with few recorded work experiences, a graph baseline reached an AUC of only 0.641, and multimodal interview scoring added negligible predictive gain over text (Spearman ρ = 0.46 vs 0.45) while visual features showed adverse impact (ratio: 0.52). The advantage of fair re-ranking narrowed as employers made more selections, and screeners conformed more often to incorrect recommendations from a bag-of-words model than from a neural model. These findings demonstrate the limitations of static offline metrics. We propose a four-level evaluation framework addressing representation validity, decision-stage stress-testing, human-AI interaction, and consequential impact. Because retrieval was incomplete and inclusion decisions were not independently re-adjudicated, this review assesses methodological practices rather than claiming an exhaustive census.

Keywords: artificial intelligence; machine learning; recommender systems; knowledge graphs; career guidance; algorithmic fairness; trustworthy evaluation.

1 Introduction

AI-supported vocational and employment decisions encompass several related but fundamentally non-equivalent computational tasks. A student-facing guidance platform classifies learner competencies, suggests developmental pathways, or recommends academic majors [4, 6, 15]; a commercial job recommender retrieves and ranks open employment vacancies for an individual job seeker [7, 11]; and a recruiter-facing applicant tracking system prioritizes and screens candidate dossiers for an organizational opening [8, 9, 10, 16]. Although these tasks are frequently grouped together under the umbrella of

'talent intelligence' or 'AI in human resources (HR)', their underlying decision-makers, optimization targets, and potential harms diverge substantially.

The critical methodological challenge in this domain lies in the divergence between what computational models predict and what vocational stakeholders genuinely care about. In typical offline machine learning pipelines, the operational training label is rarely a long-term measure of employment success, vocational fulfillment, or sustained workplace productivity. Instead, models are trained on available digital proxies: a self-reported questionnaire category [6], a set of explicit skill tokens extracted from job descriptions [3, 12], an initial vacancy click or view [10], a digital chat initiation [10], a formal job application submission [9, 11], a job-offer outcome [8], or subjective ratings of recorded video interviews [19]. None of these operational proxies is equivalent to fair, durable, and productive career development. For example, a longitudinal career-path model learns to predict historically recorded job-to-job transitions from online resume databases [1, 2], inherently reflecting historical labor market conventions, structural stratification, and economic gatekeeping rather than optimal or equitable trajectories. Similarly, an automated candidate screening model trained on historical recruiter hiring logs learns to replicate the cognitive shortcuts, prestige heuristics, and systemic demographic biases of organizational screeners [8, 9, 16].

This tension is especially pronounced in recent architectural innovations. Natural language processing (NLP) models utilize dense sentence embeddings and large language models (LLMs) to expand job vacancies with inferred implicit competencies [7]; hierarchical neural networks incorporate topic-aware attention to align resume experience sections with job requirement clauses [8]; explicit-implicit feature fusion networks combine structured tabular metadata with unstructured text [9]; sequential and state-space architectures model temporal progression and skill accumulation over ordered professional histories [1, 10]; and heterogeneous career knowledge graphs construct multi-relational networks connecting skills, educational degrees, job titles, and employers [5, 12]. While each of these architectures addresses a specific representation or ranking challenge, an improvement on an offline proxy metric (such as area under the curve (AUC), Mean Reciprocal Rank (MRR), or Recall@K) does not guarantee improved real-world career guidance or equitable hiring outcomes [14, 15]. Furthermore, algorithmic fairness interventions that equalize exposure or enforce demographic parity in top-ranked search results do not ensure fair downstream human selections, as fair rankings can shift who is selected first without equally changing the overall share of underrepresented candidates selected [14], and screeners' reliance on recommendations differs by model [16].

To establish a transparent and reproducible foundation, this review synthesizes evidence from an automated screening corpus of 181 candidate extraction studies and conducts an in-depth empirical appraisal of 19 purposively selected primary studies. Rather than pooling non-comparable accuracy scores or attempting to crown a single 'state-of-the-art' architecture, this paper addresses four core research questions (RQs):

- RQ1 (Tasks & Modalities): What computational functions, decision stages, and input modalities are represented across current career and talent AI systems?

- RQ2 (Labels & Validation): What operational targets, comparative baselines, and validation designs support reported performance claims, and how do models perform under realistic stress-testing?

- RQ3 (Taxonomies & Psychometrics): How are occupational taxonomies (in the appraised studies, the European Skills, Competences, Qualifications and Occupations taxonomy, ESCO) and psychometric constructs operationalized, and what construct validity gaps remain unaddressed?

- RQ4 (Trustworthiness & Human Decisions): Which algorithmic fairness interventions, explainability mechanisms, and joint human-AI decision outcomes are empirically verified in operational contexts?

The core contribution of this review is an explicit analytical separation between computational tasks, operational proxy labels, evaluation designs, and downstream consequential outcomes. By deconstructing the full decision pipeline across both education-to-work guidance and automated organizational recruitment, we establish a four-level evaluation ladder that enables researchers, system developers, and auditing bodies to rigorously assess the validity and trustworthiness of vocational AI systems.

2 Methods

2.1 Review design and search

This study adopted a structured methodological review design combining purposive literature retrieval with an in-depth analytical assessment of selected studies, rather than aiming to constitute an exhaustive, multi-database systematic review. Accordingly, we use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 only to structure the reporting of selection accounting, and we do not claim full PRISMA compliance or the evidentiary standing of a systematic review. Two complementary semantic searches were conducted using Elicit as a literature-search platform on 2 October 2026, searching over 138 million papers including all of Semantic Scholar and OpenAlex. The searches returned 1,000 results in total; 20 duplicates were removed, leaving 980 records for screening. Elicit was used for retrieval, title and abstract screening, full-text screening and data extraction. The verbatim search queries, screening criteria and extraction instructions are provided in ESM_1.

The first search targeted empirical applications of machine learning (ML), deep learning (DL), natural language processing (NLP), knowledge graphs (KG), and artificial intelligence (AI) in career guidance, career-path prediction, skill extraction and recommendation, person-job fit, resume-based job matching, and recruitment evaluation. The second independent search focused on empirical investigations of algorithmic fairness, demographic bias, explainability, transparency, human oversight, and trustworthy evaluation in career recommendation, algorithmic hiring, person-job matching, and automated recruitment. Fairness-related terminology was intentionally omitted as a mandatory filter in the core search query to allow technical matching and recommender studies to be identified even when their primary focus was predictive accuracy or algorithmic efficiency.

No publication-year limit was applied at search stage: 73 of the 980 records predate 2020, and the 2020-2026 limit was applied through the full-text 'eligible publication' criterion. The accompanying audit workbooks document unique record identifiers, abstract- and full-text eligibility decisions, categorized exclusion reasons, and extracted technical attributes. Screening was conducted by Elicit against criteria written by the first author at both the title/abstract and full-text stages; at title/abstract stage, records rated 'Maybe' advanced, and at full-text stage, records rated 'No' or 'Maybe' on any criterion were excluded. Although the review did not implement prospective protocol registration or dual-independent screening, its fully documented selection accounting and record-level audit workbooks provide a transparent and verifiable foundation for examining the current methodological landscape.

2.2 Selection accounting and candidate landscape

The literature selection accounting is arithmetically reconciled in Fig. 1, adapted in layout from PRISMA 2020 reporting standards [20] to provide a transparent audit trail. The initial export comprised 980 records after duplicate removal. A later title and DOI check found three works with more than one version (preprint, published or reprinted) and two proceedings volumes listed twice; these remain in the 980 as exported, and no work appears twice among the 181 studies. At the title and abstract screening stage, 265 records were excluded: 155 as not a primary empirical model report and 110 as outside an in-scope career or matching task, advancing 715 records to the full-text retrieval stage. Of these 715 sought reports, 409 (57.2%) were recorded as full text not available to the screening tool; the records do not give a reason for individual cases.

The remaining 306 publications underwent full-text assessment against three criteria: (i) eligible publication, meaning a peer-reviewed primary journal or conference paper published in 2020-2026, excluding unreviewed preprints and non-primary publications; (ii) implemented in-scope model, meaning an AI/ML or computational recommendation model for career guidance, career-path prediction, occupational or skill recommendation, job recommendation, or candidate-job matching that was implemented and evaluated; and (iii) quantitative evaluation, meaning an identified evaluation dataset or population and at least one quantitative outcome. This assessment excluded 125 reports: 57 on quantitative evaluation, 42 on eligible publication, and 26 on implemented in-scope model. This yielded a final candidate extraction set of 181 studies, all documented in the accompanying evidence workbook.

Analysis of the 181 candidate extraction studies provides a comprehensive view of the rapid evolution of this field. Fig. 2 illustrates the annual publication distribution: research activity accelerated dramatically from 2020 (12 studies; 6.6%) and 2021 (6 studies; 3.3%) to moderate levels in 2022 (17; 9.4%), 2023 (16; 8.8%), and 2024 (17; 9.4%), followed by a massive surge in 2025 (51 studies; 28.2%) and 2026 (62 studies; 34.3%). Over 62% of the candidate literature has appeared within the last two years, coinciding with the widespread adoption of large language models, dense contextual embeddings, and automated resume parsing pipelines (2026 is a partial year, so its count is a lower bound for the full year).

Fig. 3 shows the primary computational task of the 181 candidate studies: Algorithmic Fairness & Bias Analysis (58; 32.0%), Student Career Guidance (48; 26.5%), Person-Job Matching & Resume Screening (42; 23.2%), Job Recommendation (9; 5.0%), Career Path & Trajectory Prediction (9; 5.0%), Skill Extraction & Recommendation (9; 5.0%), Other AI-HR Operations (6; 3.3%). A fairness evaluation was coded as present in 61 studies (33.7%). These shares describe the retrieved corpus and are shaped by the search design, because one of the two queries targeted fairness directly; they should not be read as evidence that any topic is over- or under-researched in the field as a whole.

Fig. 4 shows the underlying model architectures across the 181 candidate studies: Classical ML (58; 32.0%), Hybrid (40; 22.1%), Transformer / LLM (26; 14.4%), Not specified / No model (e.g., audit) (25; 13.8%), Deep Learning (non-Transformer) (23; 12.7%), Recommender System (CF / MF) (5; 2.8%), Graph Neural Network / Knowledge Graph (4; 2.2%).

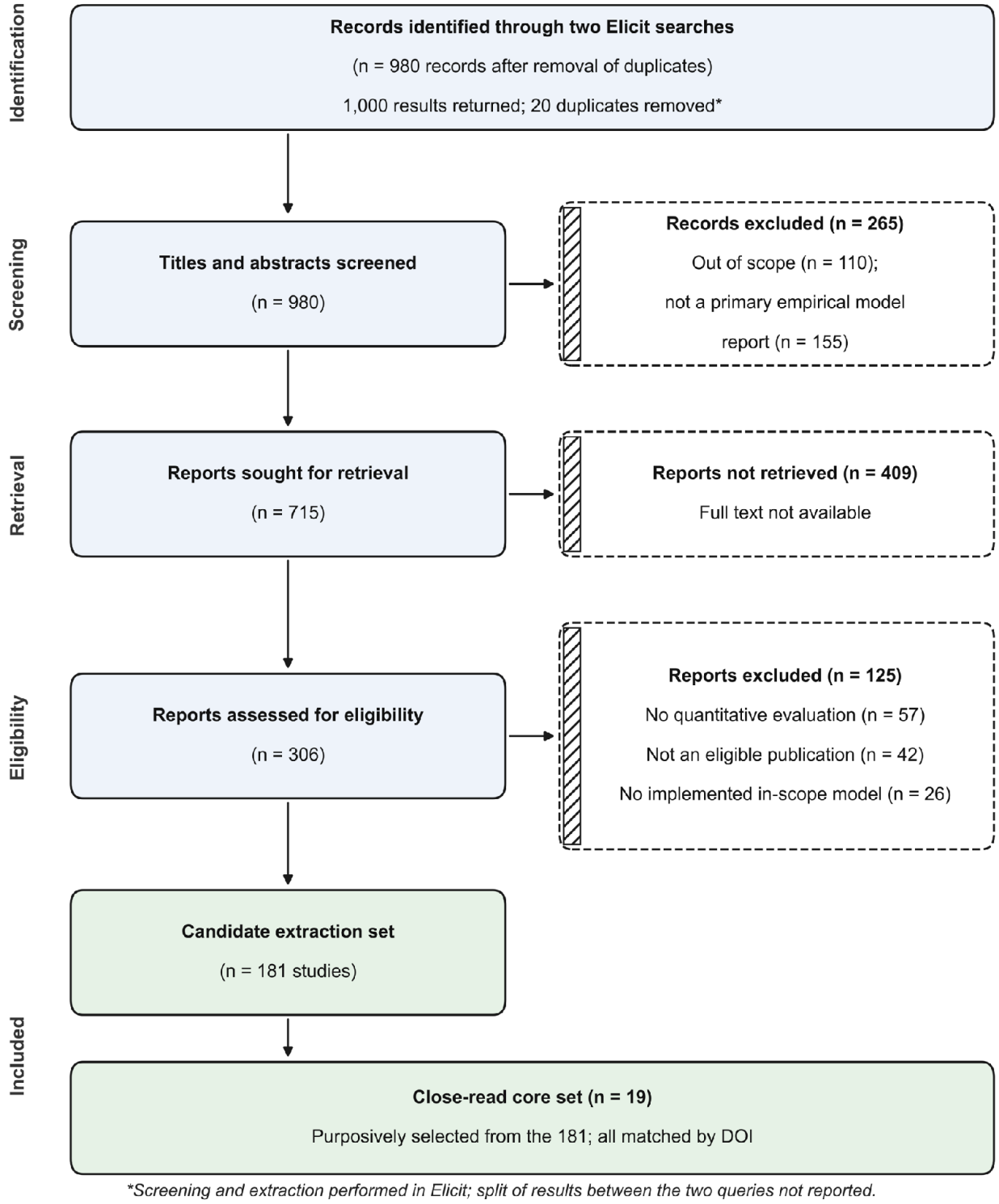

Fig. 1 Descriptive study-selection flow. Reconciled literature search, screening, and full-text eligibility accounting adapted from PRISMA 2020 guidelines. Unretrieved reports (409) are explicitly separated from substantive full-text exclusions (125). '181' denotes the candidate automated extraction set, from which 19 primary studies were purposively selected for in-depth appraisal. Layout adapted from PRISMA 2020 [20]

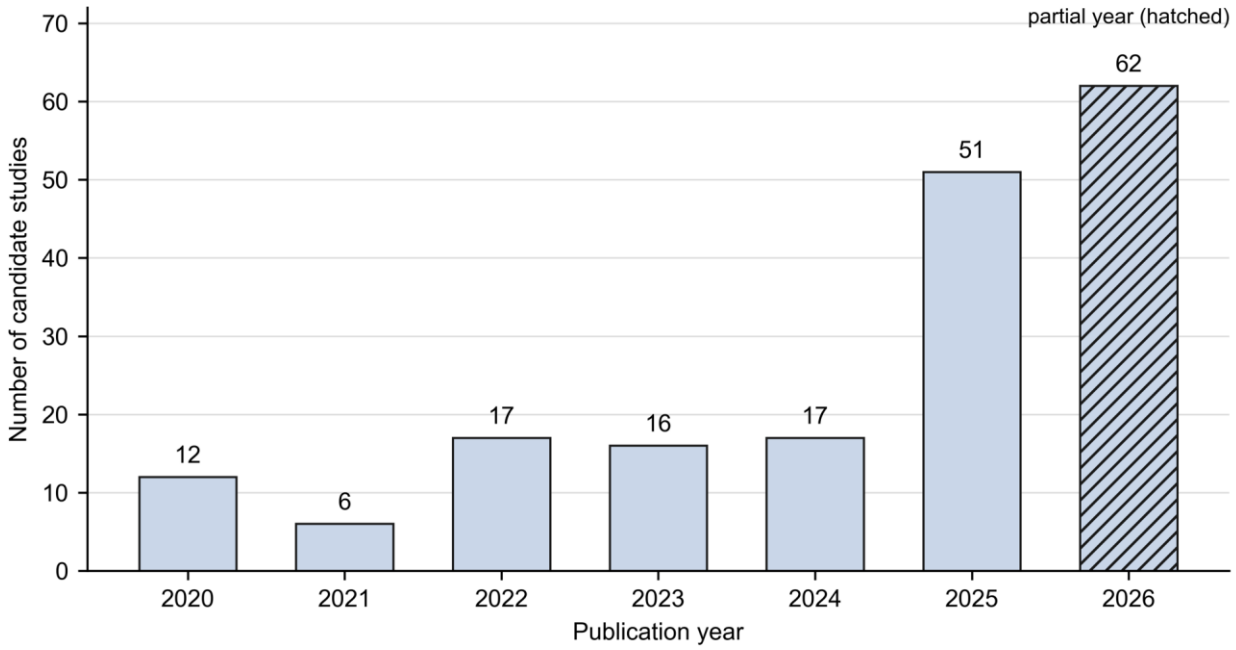

Fig. 2 Publication year distribution of candidate extraction studies (N=181). Annual distribution of the 181 candidate extraction studies published between 2020 and 2026. The field shows rapid recent expansion, with over 62% of all studies appearing in 2025 and 2026 alone (2026 is a partial year, shown hatched).

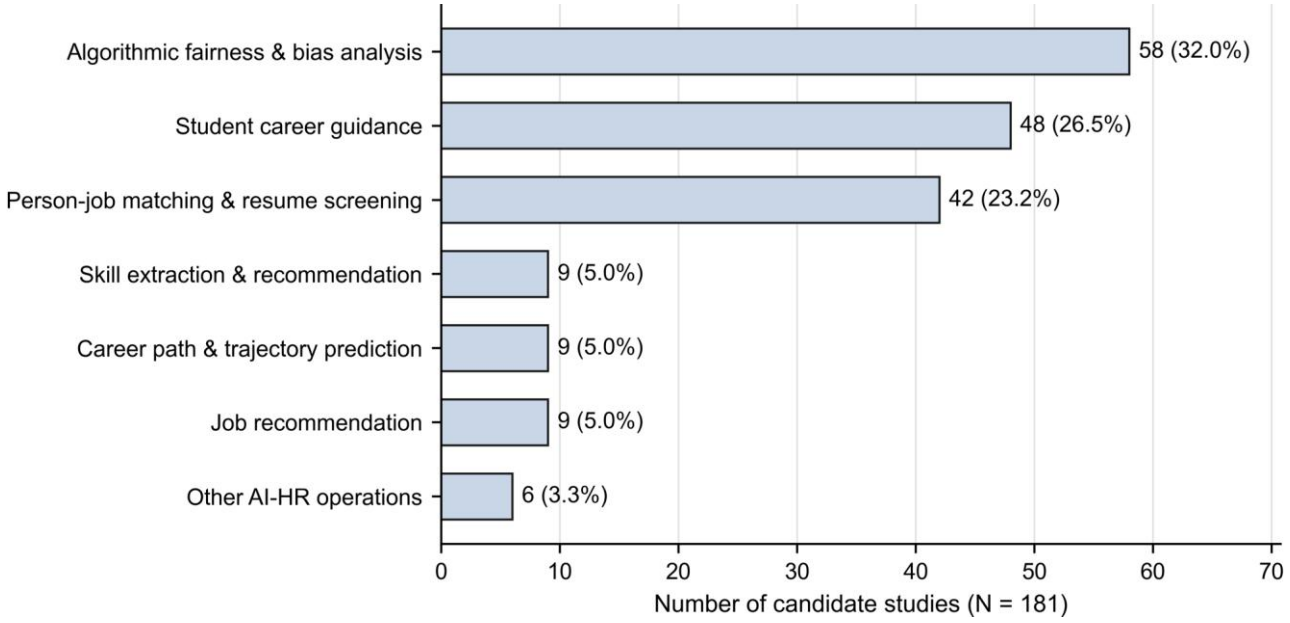

Fig. 3 Computational task breakdown across candidate extraction studies (N=181). Distribution of primary computational tasks, coded by the authors (Section 2.3). Shares describe the retrieved corpus, not the field as a whole.

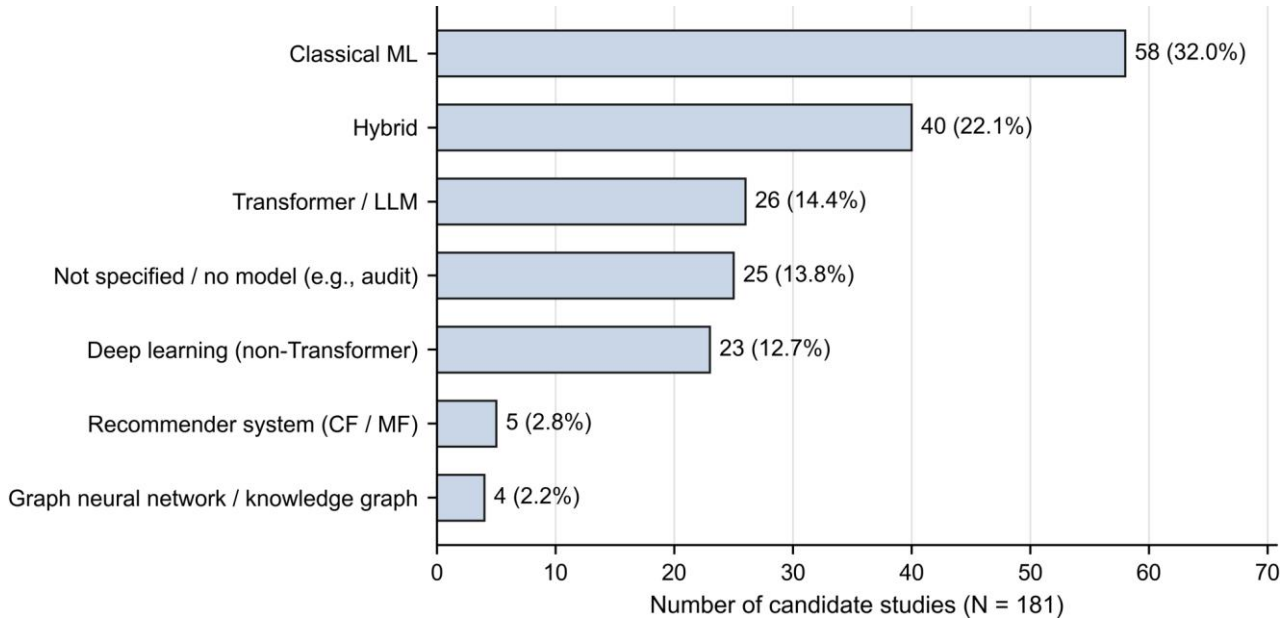

Fig. 4 Model architecture distribution across candidate extraction studies (N=181). Distribution of seven codebook architecture classes, coded by the authors (Section 2.3). CF, collaborative filtering; MF, matrix factorization.

2.3 Extraction, appraisal and synthesis

Coding used each study's title, year, venue and three Elicit-generated extraction fields, with the DOI link consulted where these were insufficient; the full papers were not systematically read. The second author independently coded a random sample of 30 studies (Python random.sample, seed 20261102). Agreement between the two authors was 86.7% for task (Cohen's kappa = 0.814), 86.7% for architecture (kappa = 0.832) and 90.0% for fairness evaluated (kappa = 0.769). Disagreements, affecting 6 of the 30 studies, were resolved by discussion, and the agreed labels were used as final. The codebook, final labels and double-coded sample are provided in ESM_2.

To move beyond high-level bibliometric distributions and investigate the empirical validity of reported claims, we purposively selected 19 primary studies for close-read critical appraisal. Rather than drawing a random sample, these 19 studies were chosen to maximize coverage across decision stages, input modalities, model architectures, dataset scales, and trustworthiness dimensions. All 19 studies correspond one-to-one by DOI to records in the candidate extraction set, and the full text of each was obtained for appraisal.

Verification of the 19 primary studies. Reported values were checked against the full text of each study, with a second check of the principal tables using page images. The first and second authors then confirmed the main reported values against the PDFs.

Fig. 5 maps these 19 primary studies across three critical dimensions: publication timeline (x-axis), computational task category (y-axis), and empirical dataset scale (bubble size, scaled logarithmically by total applicant profiles, resumes, or interaction pairs). The primary evidence base spans four orders of magnitude in dataset size, from small-scale educational samples of 215 student profiles [4] and 2,400 scraped resumes [12] to mid-scale corpora of 6,000 job descriptions [3] and 13,129 Vietnamese curricula vitae (CVs)

[5], up to large corpora of 478,752 person-job pairs [10], 1.01 million psychometric survey responses [6], 1.1 million job descriptions [7], 1.3 million resumes [9], 1.49 million applications [8], 2.2 million job ads from one French region [11], and 3.9 million Indeed profiles [1].

Fig. 6 analyzes the evaluation designs adopted across these task categories. While Person-Job Matching and Trajectory Modeling studies frequently utilize offline train/test partitions or cross-validation, they differ sharply in their temporal rigor: one study introduces a strict chronological holdout [9] and one adds a user cold-start test on public benchmark data [11], whereas others rely on random record partitioning, random holdouts or a held-out split whose method is not stated [1, 2, 3, 4, 6, 7, 8, 10]. Each primary study is assigned in Fig. 6 to its single primary validation design (n = 19); two studies fall outside the standard offline designs, one using an expert-rated benchmark [5] and one using unsupervised semantic evaluation [12]. Conversely, studies focusing on Algorithmic Fairness and Teamwork are the only category incorporating controlled human-subject choice experiments [14, 15, 16] or platform audits, one cooperative [17] and one external [18], highlighting that evaluation methodology is strongly bifurcated between offline metric optimization and socio-technical assessment. A representativeness caveat applies: 15 of the 19 primary studies (79%) were published in 2020-2023, whereas 62.4% of the candidate set appeared in 2025-2026, so the close-read appraisal captures the pre-LLM literature better than the most recent wave.

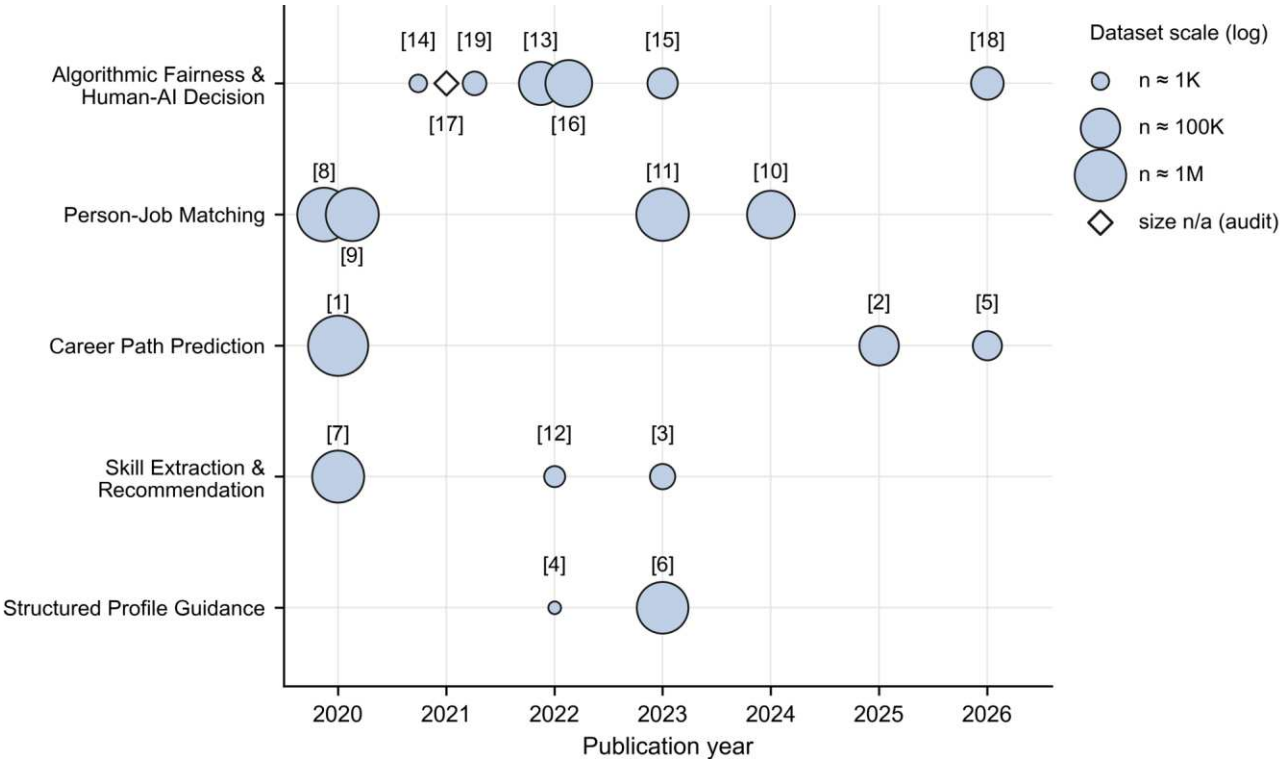

Fig. 5 Multi-dimensional landscape of the 19 primary close-read studies. Mapping the 19 primary close-read studies across publication timeline (x-axis), computational task category (y-axis), and empirical dataset scale (bubble size, log-scaled by total profiles, applications, or interaction pairs). Studies span four orders of magnitude in dataset size; a

diamond marks an audit with no dataset size.

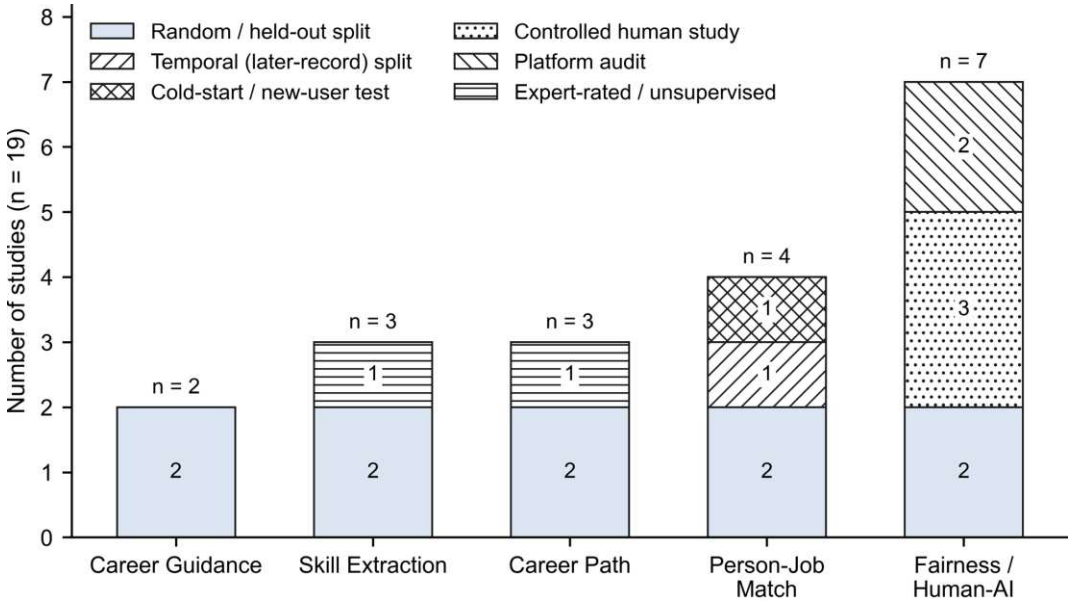

Fig. 6 Validation design patterns across computational task categories. Distribution of primary empirical validation designs across task categories in the primary evidence set (n = 19; each study is counted once). Technical matching tasks rely on offline random splits, temporal holdouts or cold-start tests, whereas human-subject choice trials and platform audits are restricted to fairness and human-AI studies; each design has its own fill pattern.

3 Results

3.1 Computational tasks and model families

The computational landscape of AI in career guidance and talent matching is structured along four core dimensions: the primary vocational task, the input modalities processed, the underlying algorithmic architecture, and the operational evaluation target. Fig. 7 provides a task-first taxonomy mapping input representations through method families to computational functions. This taxonomy clarifies that different model families do not represent competing answers to a single question; rather, each family is specialized to solve a specific functional problem within the broader vocational pipeline.

Table 1 provides a systematic comparison of six major AI approach families identified across the examined studies, together with fairness and debiasing interventions, detailing what each learns from, its representative task, its computational strengths, and its fundamental methodological limitations. Classical machine learning methods excel at tabular, interpretable classification on structured records [4, 6] but cannot handle unstructured narrative text or temporal transitions. Dense text encoders and NLP architectures map vacancies and candidate bios into semantic vector spaces to identify implicit competencies [7, 12] but evaluate relevance through text cosine similarity rather than candidate capability. Neural person-job matching architectures learn complex interaction patterns between resume sections and job descriptions [8, 9, 10] but optimize against historical recruiter behavior that encodes systemic organizational biases. Sequential and state-space models capture career trajectories and skill growth over time [1, 2, 10] but treat recorded transitions as normative paths without observing unchosen

counterfactual alternatives. Career knowledge graphs structure explicit multi-hop relations between entities [5, 12] but remain vulnerable to cold-start sparsity and vocabulary drift. Finally, large language models provide impressive few-shot text reasoning [2, 5] but can suffer from high inference latency [5], overfitting under small-sample fine-tuning [2] and hallucination risk [5].

A critical finding of this review is that offline performance metrics derive their entire meaning from the operational label used as ground truth. Table 2 deconstructs six prevalent operational labels, contrasting what each validly measures against the critical validity gaps that remain untested. A high classification accuracy on a psychometric dataset merely reflects statistical consistency across survey items; it does not establish that recommending a matching occupation will yield vocational success or life satisfaction [6]. Similarly, optimizing click-through or recruiter-chat probabilities in commercial platforms models surface platform engagement rather than candidate qualification, job offer conversion, or sustained employment [10, 11].

Table 3 categorizes the six primary validation designs observed in the literature, specifying the exact transfer question each design answers and the critical real-world questions it leaves unaddressed. Random cross-validation evaluates within-distribution interpolation but cannot assess temporal generalizability or performance on emerging occupations [1, 4]. Strict chronological holdouts test whether a model predicts future application cycles without look-ahead bias [9], but ignore macroeconomic shifts and new-firm cold starts. Controlled human choice trials measure how human decision-makers respond to algorithmic recommendations [14, 15, 16], yet laboratory trials may not replicate organizational hiring committee behavior.

Fig. 8 synthesizes these components into an integrated career and talent decision pipeline. The diagram delineates the multi-stage progression from raw candidate input and multi-modal representation through algorithmic scoring and ranked recommendation, culminating in the human selection decision and downstream career consequence. Critically, Fig. 8 establishes three continuous audit pillars that must span the entire pipeline: (i) Representation & Taxonomy Validity; (ii) Offline Robustness & Stress-Testing; and (iii) Point-of-Impact Fairness & Human Oversight.

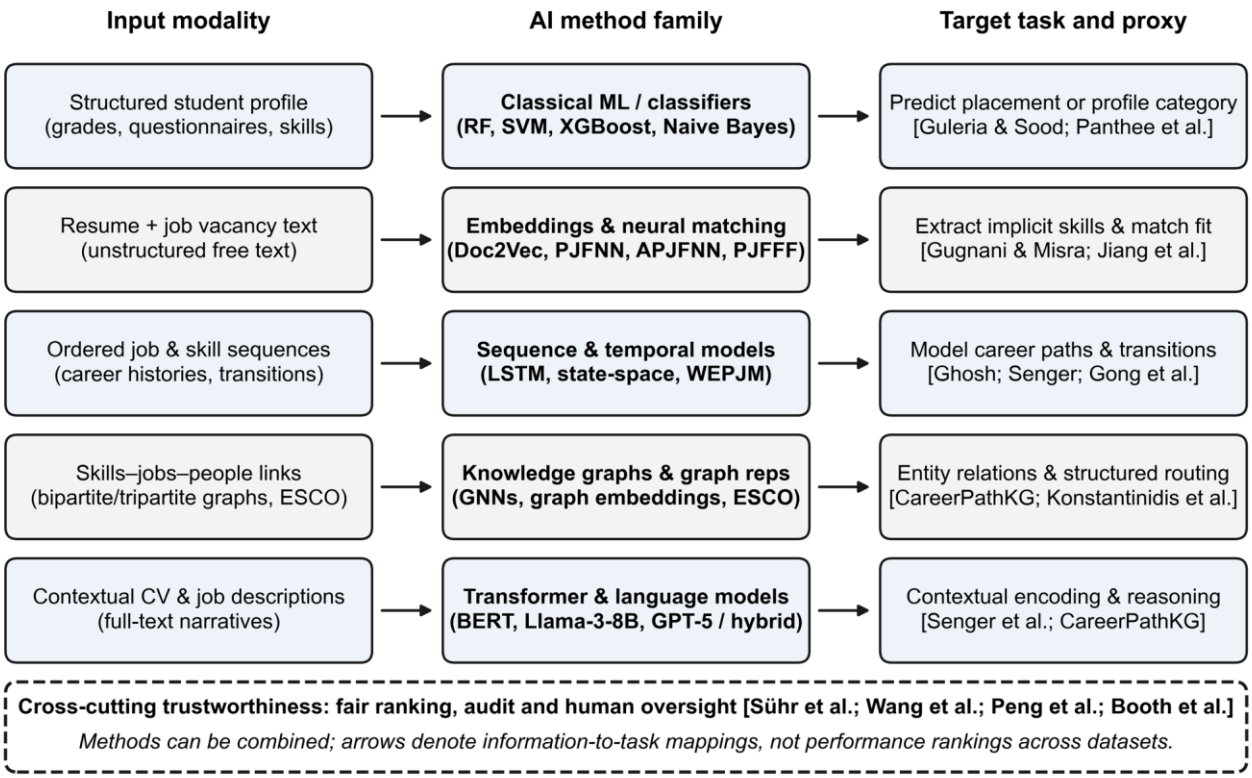

Fig. 7 Task-first taxonomy of AI approaches in career guidance and talent matching. Mapping input representations through computational method families to specific vocational tasks. The framework highlights that model families solve distinct representation and ranking problems rather than competing for a single objective.

Table 1 AI approaches in the examined career-guidance and talent-matching studies

AI Approach	Learned Input Modalities	Representative Task in Corpus	Computational Advantages	Core Methodological / Validity Limitation
Classical ML / Structured Classifiers	Tabular applicant attributes, psychometric questionnaires, academic grades	Placement outcome prediction; personality and learning-style typing [4, 6]	Interpretable feature attributions (decision trees, linear models); robust baselines on tabular data	Predicted academic placement or psychometric category does not validate long-term vocational suitability
Text Embeddings / NLP Encoders	Free-text job descriptions, resume experience paragraphs, vacancy titles	Implicit skill extraction; vacancy retrieval and recommendation [7, 12]	Captures latent semantic relatedness beyond exact keyword matching; models implicit competencies	Extracted skills reflect job ad phrasing rather than candidate mastery; offline retrieval ignores hiring decisions
Neural Person-Job Matching	Bi-directional resume text and job requirement text; historical interaction logs	Candidate-job fit scoring; job-offer outcome prediction [8, 9, 10]	Models fine-grained token-level cross-attention, hierarchical ability matching, and explicit-implicit feature fusion	Platform interaction proxies (views, recruiter chat, recruiter and offer decisions) reflect historical institutional gatekeeping
Sequential & Temporal Models	Ordered career roles, chronologically logged promotions, skill development timelines	Career trajectory prediction; path-aware candidate matching [1, 2, 10]	Encodes temporal progression, tenure dynamics, and experience momentum rather than static point-in-time snapshots	Historical career transitions reflect legacy labor-market inequalities and unobserved counterfactual paths
Graph & Knowledge Base Architectures	Multi-relational graphs of skills, occupations, educational degrees, and company entities	Graph-based talent matching; career navigation intelligence [5, 12]	Explicitly structures multi-hop ontological relations; provides interpretable relational audit trails	Susceptible to graph cold-start disconnectedness; requires extensive curation of regional taxonomy drift

AI Approach	Learned Input Modalities	Representative Task in Corpus	Computational Advantages	Core Methodological / Validity Limitation
Transformer & Large Language Models	Contextual natural language resumes, comprehensive vacancy specifications, narrative bios	End-to-end resume evaluation; zero-shot career guidance [2, 5]	Rich contextual understanding; few-shot and zero-shot reasoning over diverse professional narratives	Inference latency; hallucination risk; performance degradation under fine-tuning on small datasets
Algorithmic Fairness & Debiasing Interventions	Candidate ranking slates, protected demographic attributes (gender, age, ethnicity), decision exposure	Fair talent search ranking; debiased major recommendation; team decision audits [13, 14, 15, 17]	Enforces demographic parity, exposure balance, or disparate impact thresholds (>0.80) at model output	Fairness at the scoring or ranking stage does not guarantee equitable human selection decisions due to employer behavioral biases

Table 2 What operational labels in career and talent AI actually measure

Operational Label / Proxy	Decision Stage & Stakeholder	What the Model Validly Measures	Critical Untested Validity Gaps
Questionnaire category / placement status	Student Career Guidance (Individual Learner)	Statistical correlation with existing survey labels or historical campus hiring events	Does not demonstrate personalized counseling efficacy, career longevity, or vocational flourishing
Extracted skill tokens / ESCO concepts	Skill Intelligence & Profiling (Job Seeker / System)	Textual alignment with vacancy phrasing or occupational taxonomy definitions	Does not verify authentic candidate skill mastery, depth of expertise, or job-market value
Platform clicks / job views / recruiter chat initiation	Candidate Sourcing & Job Browsing (Platform / Recruiter)	Likelihood of initial user engagement and surface browsing behavior	Does not predict candidate application completion, qualification match, or durable placement

Operational Label / Proxy	Decision Stage & Stakeholder	What the Model Validly Measures	Critical Untested Validity Gaps
Recruiter decision / job-offer outcome	Resume Screening (Employer / HR Screener)	Conformity to historical recruiter screening habits and organizational prestige heuristics	Conflates employer screening bias with genuine worker productivity; fails to measure counterfactual quality
Simulated video hireability rating	Asynchronous Interview Assessment (Hiring Panel)	Statistical alignment with human raters' subjective impressions of recorded video clips	Rater impressions correlate weakly with on-the-job performance while introducing severe acoustic/visual demographic bias
Ranked slate exposure / demographic parity metric	Candidate Search Ranking (Recruiter / Auditor)	Mathematical balance of group representations within top-k recommendation lists	Fails to ensure minority selection: human recruiters frequently skip fair-ranked minority candidates in lower choices

Table 3 Evaluation design and the specific validity question it answers

Validation Design	Representative Studies in Corpus	Specific Validity Question Answered	Critical Transfer Question Left Unanswered
Random or Unspecified Holdout / Cross-Validation	Ghosh et al. [1]; Senger et al. [2]; Guleria & Sood [4]; Panthee et al. [6]; Qin et al. [8]; Gong et al. [10] (split method not stated); Booth et al. [19]	Can the model accurately interpolate labels within identically distributed, in-sample candidate profiles?	Fails to test temporal drift, out-of-domain generalizability, emerging occupations, or new labor market entrants
Strict Chronological / Temporal Holdout	Jiang et al. [9]	Does the model maintain predictive power on future application cycles without future-information leakage?	Does not assess new-firm cold start, macroeconomic labor shifts, or whether recommendations optimize worker welfare

Validation Design	Representative Studies in Corpus	Specific Validity Question Answered	Critical Transfer Question Left Unanswered
Explicit User / Job Cold-Start Evaluation	Bied et al. [11] (on RecSys 2017 data)	How severely does performance degrade when recommending for first-time job seekers or novel vacancies without interaction history?	Does not measure long-term placement quality or wage sustainability for marginalized, low-qualification workers
Controlled Human-in-the-Loop Choice Trials	Sühr et al. [14]; Wang et al. [15]; Peng et al. [16]	How do human screeners and advisees alter their final selection decisions when interacting with algorithmic outputs?	Does not verify whether lab/crowdsource trial behaviors translate to organizational hiring panels under commercial incentives
External or Cooperative Platform Audit	Wilson et al. [17]; Behzad et al. [18]	Does the commercial proprietary algorithm exhibit systematic exposure disparities or demographic bias across user subgroups?	Cannot observe closed-door hiring interviews, candidate offer acceptance, or institutional onboarding decisions
Expert-Rated Benchmark / Unsupervised Evaluation	Le et al. [5]; Konstantinidis et al. [12]	Does the system agree with expert judgments or produce coherent semantic structure in the absence of observed outcome labels?	Small expert panels or unsupervised similarity cannot establish predictive validity against hiring or career outcomes

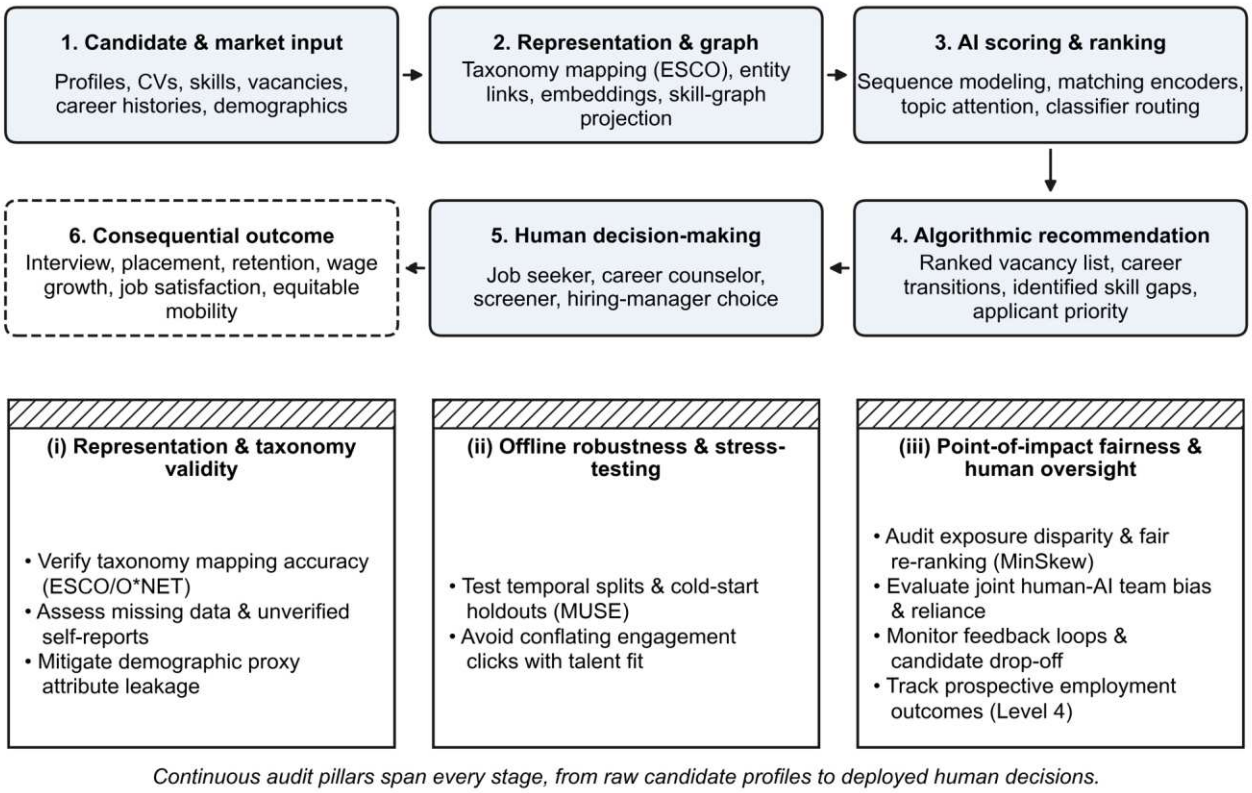

Fig. 8 End-to-end AI career/talent decision pipeline and continuous audit pillars. Two-tier operational decision flow mapping candidate input through multi-modal representation, algorithmic scoring, ranked recommendation, human selection, and consequential impact. Three continuous audit pillars span the pipeline to ensure representation validity, offline robustness, and point-of-impact fairness.

3.2 Structured profiles, skill taxonomies, and career trajectories (RQ1-RQ3)

Structured-profile systems provide a baseline for vocational AI by mapping academic marks, demographic attributes, and psychometric inventories directly to classification targets. Panthee et al. [6] trained supervised classifiers on 1,015,342 Big Five questionnaire responses from a Kaggle dataset to predict Big Five personality traits, and an auxiliary dataset to classify VAK (Visual, Auditory, Kinesthetic) learning styles. Random Forest achieved 91.14% accuracy on Big Five traits (substantially outperforming Decision Trees at 78.25%), while Support Vector Machines (SVM) achieved 90.50% accuracy on VAK styles (outperforming XGBoost at 88.42% and Random Forest at 83.82%). In a parallel educational setting, Guleria and Sood [4] evaluated supervised classifiers on a campus placement dataset of 215 students with academic and employability attributes using 10-fold cross-validation. Naive Bayes achieved the highest predictive performance (Accuracy = 91.16%, Recall = 0.912, F-measure = 0.907), while Random Forest yielded 85.58% and J48 Decision Trees 82.79% accuracy. While Guleria and Sood emphasize white-box interpretability for student counseling, neither study demonstrates that classifying an existing questionnaire category or predicting historical campus placement improves counseling efficacy, long-term career satisfaction, or vocational flourishing.

Skill-centric architectures operationalize guidance as the extraction and recommendation of competence tokens. Ong and Lim [3] developed SkillRec, scraping 6,000 job listings

across three employment platforms over six months and building a standardized vocabulary of 589 distinct skills. When predicting required skills from job titles, Bidirectional Encoder Representations from Transformers (BERT) coupled with a neural network (NN) achieved an apparent overall accuracy of 98.70% (compared to 97.28% for FastText+NN and 47.68% for Bag-of-Words, BoW). However, the F1-score was only 0.4973 (FastText+NN: 0.4931; BoW: 0.2887). We interpret this divergence between 98.7% accuracy and 0.497 F1 as consistent with class imbalance across the skill vocabulary; the paper itself does not report per-skill results. Konstantinidis et al. [12] bypassed supervised training by developing an unsupervised pipeline combining Siamese BERT sentence embeddings, a Faiss similarity index, and the ESCO taxonomy. Evaluated across 2,400 Kaggle resume PDFs, their system extracted skill entities and constructed a bipartite candidate-skill graph. However, because evaluation relied on unsupervised semantic similarity (Mean Average Precision, MAP, and Average Discounted Cumulative Gain, ADCG) without ground-truth recruiter validation, taxonomy mapping errors propagate uncorrected into downstream matching. With respect to RQ3, ESCO is the only occupational taxonomy used operationally among the 19 primary studies [2, 12], none of which reports mapping error rates, and psychometric constructs appear solely as classification labels [6]; evidence on other taxonomies such as the Occupational Information Network (O*NET) is therefore outside the scope of this appraisal.

Sequential trajectory models capture the temporal dynamics of professional careers. Ghosh et al. [1] modeled career progression using an interpretable Monotonic Nonlinear State-Space (MNSS) model evaluated on LinkedIn (1,136,231 users) and Indeed (3,945,040 users) datasets under 5-fold cross-validation. By constraining latent skill mastery to be non-decreasing over time, MNSS achieved a skill Mean Percentile Rank (MPR) of 2.81 ± 0.01 (compared to 2.94 ± 0.02 for an unconstrained nonlinear state-space model, NSS) and Precision@100 of 67.69 ± 0.04% on LinkedIn data, alongside company transition MPR of 4.75 ± 0.01. While MNSS successfully plans feasible career trajectories, the authors explicitly noted that the model learns solely from successful, observed career paths; it cannot observe counterfactual rejections, unobtained credentials, or structural barriers that prevent candidates from pursuing theoretically optimal paths.

Senger et al. [2] conducted an extensive comparative benchmark across five dataset configurations, evaluating Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), Linear, and Llama-3-8B-Instruct models on next-occupation prediction across 3,039 ESCO occupations. Using official 80/10/10 splits on the DECORTE corpus (2,482 English resumes) and KARRIEREWEGE+ (100,000 German employment-agency resumes), their findings challenge several prevailing assumptions. First, fine-tuning dense text representations (all-mpnet-base-v2) substantially boosted MLP performance by up to 10 percentage points (achieving MRR = 25.57, Recall@5 = 35.13%, Recall@10 = 42.56% on DECORTE), whereas LSTMs achieved lower performance (MRR = 15.25). On the German profiles, the MLP reached MRR = 43.58 on K+occ (Linear: 40.56), while on K+ESCO Linear and MLP were nearly equal (47.75 and 47.72) and LSTM reached 49.13. Using only job titles instead of titles and descriptions reduced Linear MRR on K+ESCO from 47.75 to 38.87. Senger et al. also examined synthetic data as an error source: in manually reviewed incorrect predictions, the mean correctness score of synthetic text fell

from 4.72 to 3.76 on K+occ and from 4.82 to 4.02 on K+cp. Furthermore, fine-tuning Llama-3 for one epoch on the smaller DECORTE corpus reduced MRR from 18.2 (vanilla) to 13.96.

Le et al. [5] bridged trajectory modeling and candidate evaluation by developing CareerPathKG, a directed career knowledge graph constructed from 13,129 CVs and 7,322 job descriptions (JDs) across 13 technology-related job categories, with an average of 15.10 skills per node. Evaluated against 100 anonymized CVs assessed by five HR experts, CareerPathKG achieved an assessment root mean square error (RMSE) of 8.9 (lower than commercial LLMs: GPT-5 at 12.4, Gemini-2.5-Pro at 11.7, and Claude 4.5 at 12.1; no significance test reported) and a CV-JD matching Recall@10 of 78.3% (vs GPT-5 at 68.9% and Gemini at 70.4%). Furthermore, the graph-guided system averaged 0.24 seconds per CV-JD matching query, compared to 13.9 seconds for GPT-5 and 22.4 seconds for Gemini-2.5-Pro. However, qualitative error analysis revealed that CareerPathKG suffered from semantic over-matching, generating false-positive recommendations when candidate CVs shared superficial keywords with job requirements (e.g., conflating 'data pipeline processing' with 'CI/CD deployment pipelines').

3.3 Candidate-job matching and recruitment platforms (RQ1-RQ2)

Candidate-job matching architectures differ fundamentally in their operational definition of a 'match'. Gugnani and Misra [7] addressed vocabulary mismatch by training Doc2Vec embeddings on 1.1 million web-mined job postings across 52 domains to retrieve similar vacancies and infer unstated implicit skills. They report an MRR of 0.88 for matching resumes and job descriptions, a 29.4% improvement over an explicit-skills baseline, and accuracy at the first recommendation of 0.84 on 200 candidate profiles and 10,000 job descriptions. Skill extraction on 275 job descriptions achieved Precision = 0.80, Recall = 0.93, and F1 = 0.86. In our reading, inferring skills from similar postings also risks reinforcing existing occupational norms; the paper does not evaluate this risk.

Qin et al. [8] addressed talent sourcing using TAPJFNN, an end-to-end neural network incorporating hierarchical topic-based ability-aware attention to align candidate experiences with job requirements. Evaluated on a Baidu dataset of 1,495,166 applications (20,184 successful and 1,474,982 failed, where success means receiving a job offer) across 5,804 job postings and 696,251 resumes, under a random 80/10/10 split, TAPJFNN achieved AUC = 0.9300, Accuracy = 85.08%, and F1 = 84.54%. This exceeded baselines including gradient-boosted decision trees (GBDT; AUC = 0.8551), PJFNN (AUC = 0.8729), and APJFNN (AUC = 0.8959). However, to handle extreme class imbalance (about 1.3% successful applications), negative instances were under-sampled to equal the positives for each job posting, so these metrics do not reflect real-world base rates.

Jiang et al. [9] developed PJFFF, a neural feature-fusion architecture evaluated on a real-world enterprise recruitment platform dataset (Ping An) comprising 1.3 million resumes, 13,000 jobs, and 580,000 candidates. PJFFF combines DeepFM to learn non-linear feature interactions from structured profile entities with a Text-CNN to process unstructured text. Crucially, Jiang et al. implemented a strict chronological evaluation design: 1.1 million historical applications were used for training, 100,000 for validation, and the final 100,000 chronological applications were reserved for testing, preventing temporal

information leakage. PJFFF achieved state-of-the-art performance across all metrics: AUC = 0.953, Accuracy = 92.9%, F1 = 77.1%, and Precision@Recall=0.8 = 73.3%, surpassing LightGBM (AUC = 0.917, Acc = 90.2%, F1 = 67.6%) and standard neural baselines. A graph neural network baseline (LightGCN) performed worst, with an AUC of only 0.641 and an F1 of 36.5%. The authors attribute this mainly to LightGCN not using resume and job content, and note its cold-start issue for new candidates and job posts; no dedicated cold-start evaluation was run.

Gong et al. [10] explicitly incorporated career progression into person-job fit via WEPJM, a multi-task learning framework trained on 478,752 person-job pairs (119,031 positive and 359,721 negative) from an online recruiting platform. A positive label was defined as a chat between a job seeker and recruiter, and a negative instance as viewing a job profile without further action. WEPJM combines the matching objective with three auxiliary objectives: career path identification (CPI), career path reconstruction (CPR), and career-job contrastive learning (CL). WEPJM achieved AUC = 0.929, Accuracy = 85.0%, and F1 = 83.7%, outperforming PJFNN (AUC = 0.836) and BERT with task-adaptive pre-training (BERT+TAPT; AUC = 0.887). In ablation experiments, removing all three auxiliary tasks reduced AUC from 0.929 to 0.889. Gong et al. also tested job seekers with fewer than three work experiences: performance degraded with fewer experiences, but the authors report a significant gain over text-matching baselines remained.

Bied et al. [11] evaluated public employment recommendation using proprietary data from one French region of the public employment service (Pôle emploi), encompassing 1.2 million job seekers, 2.2 million job postings, and 1.29 million applications. Their proposed MUSE.2 architecture employs a two-tier design: a coarse filtering model (MUSE.0) retrieves the top 1,000 candidate ads, followed by a multi-head neural re-ranker (MUSE.1/2) that jointly optimizes hiring and application heads. Evaluated on the Pôle emploi data with training on 85% of uniformly selected weeks and testing on the remaining weeks, MUSE.2 achieved Recall@10 = 30.1% and Recall@20 = 40.2%, significantly outperforming the RecSys challenge-winning XGBoost baseline (Recall@10 = 26.83%; Recall@20 = 35.59%). A separate user cold-start test was run on the public RecSys 2017 data, not on Pôle emploi data: Recall@100 for MUSE.2 moved from 24.9% (warm-start) to 24.0% (cold-start), while DropoutNet fell from 41.2% to 23.1%. Standalone module tests gave Recall@100 of about 15% for geographic features, about 34% for the general text module and about 4% for the skill module. Recall@10 was 31% for women and 29% for men, a statistically significant gap; this is a performance measure, not an exposure measure.

3.4 Trustworthy evaluation: fairness, explanations, and human-in-the-loop decisions (RQ4)

Empirical evaluations of algorithmic fairness, explainability, and human-AI interaction expose critical limitations in relying solely on offline model-level metrics. Delecraz et al. [13] report fairness monitoring of a job matching algorithm at Gojob, a French temporary staffing agency, using application records in the order of 200,000 rows and twofold cross-validation. In the dataset itself, Disparate Impact was close to 1 for gender (0.98) and age bands (0.87 to 1.06) but lower for residence-permit requirement (0.72), place of birth (0.78), education (0.82) and nationality (0.66). For the model, True Positive Rate (TPR)

Parity gaps ranged from 0.25 to 4.54 across sensitive attributes, against a threshold the authors describe as chosen somewhat arbitrarily. Because the operational label was a binary recruitment event, the study could not observe job duration, contract stability, or wage equity.

Wilson et al. [17] report a cooperative audit of Pymetrics, a commercial candidate screening platform, in which outside auditors worked with co-authors from the company. The platform uses 12 games and Support Vector Machines over 64 features. The audit examined whether its process enforces the four-fifths rule (minimum bias ratio above 0.8) and found that it does. However, the auditors could not directly audit the composition of the bias group used for adverse impact testing: it is drawn from players who opted in to a demographic survey and is not necessarily representative of potential job seekers. They note that if it is not representative, the bias assessments may be flawed.

The assumption that fair algorithmic rankings translate directly into fair hiring decisions was tested in the randomized experiment of Sühr et al. [14]. With 1,079 proxy employers on Amazon Mechanical Turk, participants selected four candidates from 10-person slates (7 men, 3 women) across three TaskRabbit task categories (Shopping, Moving Assistance, Event Staffing). Comparing the platform ranking (RabbitRanking) with FairDet-Greedy re-ranking, fair re-ranking increased the share of female candidates among first selections (e.g., in Moving Assistance with candidate set D1, 4.8% vs 12.3%; the largest increase across sets was 17.35 percentage points, also in Moving Assistance). However, the gap between the two rankings decreased as the number of selections increased: in Moving Assistance across the three sets, it narrowed from 17.4 percentage points at the first selection to 5.2 at four selections (our calculation from their Table 3). The authors conclude that FairDet-Greedy mainly moves female candidates higher in participants' priority order and has less impact on the overall fraction of female candidates selected.

Behzad et al. [18] conducted an independent third-party audit of LinkedIn Talent Search using Recruiter Lite, collecting rankings of around 26,000 candidates across 78 queries from the state of New York. LinkedIn's published evaluation reports that average MinSkew@100 improves from -0.259 to -0.011 under Det-Greedy. In the audit, skew for women tended to be sharply negative at early ranks, and mixed-effects tests rejected an expected MinSkew@k of -0.011 at each of k = 25, 50, 75 and 100 ($p < 0.001$), with disparities most severe at the top of the list. The audit also studied day-to-day churn in rankings, finding statistically significant group differences at k = 25 and k = 50.

Wang et al. [15] observed an even more striking behavioral reversal in student guidance. Developing a Neural Collaborative Filtering (NCF) recommender on Facebook user data (16,619 users, 1,380 academic majors, 3.5 million likes), they applied linear projection debiasing to eliminate gender bias. While offline metrics confirmed that debiased NCF successfully eliminated gender disparities in recommended majors, an online user study with more than 200 college students produced a different result: mean acceptance was significantly higher for gender-aware (biased) recommendations (0.372) than for debiased recommendations (0.279; $p < 0.01$). Perceived gender disparity was a determining factor in acceptance, suggesting that mathematical debiasing without attention to human bias may fail in practice.

Peng et al. [16] investigated human-AI teamwork in candidate screening using 397,907 Common Crawl biographies and six occupations forming three commonly confused pairs (attorney/paralegal, surgeon/physician, professor/teacher). Comparing human screeners alone with human-AI teams paired with a deep neural network (DNN), a bag-of-words (BOW) model or a random baseline, they found that higher-performing models consistently produced higher-performing hybrid teams, but the effect on bias depended on the model. When the model was wrong, the hybrid decision matched it in 70.4% to 80.9% of cases with BOW, against 50.4% to 66.3% with DNN and 56.1% to 67.3% with the random baseline. Teams with DNN became unbiased, while BOW teams appeared to reflect both human and model biases. The authors offer BOW's greater interpretability as a possible explanation, not an established cause, and caution that using a less interpretable model is not a shortcut to true bias mitigation.

Finally, Booth et al. [19] provided informative empirical evidence regarding multimodal candidate assessment. Evaluating 4,255 mock interview recordings from 733 participants (524 gender-matched subset) using Random Forest regressors across verbal (transcript), paraverbal (acoustic), and visual (facial video) features under participant-independent 5-fold cross-validation over 100 trials, they measured predictive accuracy against hireability ratings from trained raters. Adding acoustic and video modalities yielded a negligible accuracy improvement over verbal transcripts alone (Spearman $\rho = 0.45$ for verbal alone vs $\rho = 0.46$ for multimodal combined). However, gender predictability was higher for non-verbal and combined features: the area under the receiver operating characteristic curve (AUROC) was 0.73 for verbal, 0.79 for visual, 0.99 for paraverbal and 0.98 for combined. The Adverse Impact Ratio (AIR) fell to 0.52 for visual features and 0.76 for combined features (below the 0.80 four-fifths benchmark), against 0.87 for verbal. In this mock-interview sample, non-verbal sensing thus increased demographic bias with only a negligible increment in predictive accuracy.

3.5 Cross-study synthesis and empirical benchmark comparisons

To synthesize the empirical findings across the reviewed literature, Table S1 in ESM_2 presents a master synthesis of all 19 primary close-read studies across eight dimensions: bibliographic reference, decision stage and task, input modality and operational label, model architecture, dataset identity and scale, validation split design, key reported empirical metrics, and evaluation validity level. Table S2 in ESM_2 maps each study to its extraction record.

Building upon this synthesis, Table 4 extracts the quantitative benchmark and stress-test evidence from studies that conducted competitive baseline comparisons and empirical failure-mode stress tests. Table 4 directly contrasts proposed architectures against their strongest competitive baselines across eight critical vocational tasks, documenting the performance change or fairness result reported for each. The benchmark evidence indicates that strong results under standard offline testing can coexist with weaknesses under specific conditions: graph baselines without content features (Jiang et al. [9]), cold-start comparisons on benchmark data (Bied et al. [11]), few recorded work experiences (Gong et al. [10]), reduced input text and small-sample fine-tuning (Senger et al. [2]), multi-item human selection (Sühr et al. [14]), model-specific human conformity (Peng et al. [16]), and multimodal sensing (Booth et al. [19]).

Table 4 Quantitative benchmark and stress-test evidence across competitive models

Study Reference & Decision Task	Proposed Architecture	Best Competitive Baseline	Baseline vs Proposed Metrics	Empirical Stress Test / Failure Mode Evaluated	Observed Performance Degradation / Fairness Penalty
Gong et al. [10] Person-Job Matching	WEPJM (Multi-task Career-Path Consistency)	BERT+TAPT (Pretrained contextual encoder)	WEPJM AUC = 0.929, Acc = 85.0%, F1 = 83.7% vs BERT+TAPT AUC = 0.887, Acc = 80.4%, F1 = 79.8%	Fresh job seekers (fewer than 3 work experiences)	Performance degraded with fewer experiences but stayed above text-matching baselines; removing all auxiliary tasks reduced AUC from 0.929 to 0.889
Jiang et al. [9] Talent Screening	PJFFF (Explicit-Implicit Feature Fusion)	LightGBM (Gradient-boosted decision trees)	PJFFF AUC = 0.953, Acc = 92.9%, F1 = 77.1% vs LightGBM AUC = 0.917, Acc = 90.2%, F1 = 67.6%	Graph neural network baseline (no dedicated cold-start test)	LightGCN AUC = 0.641 and F1 = 36.5%; authors attribute this to not using content, and note its cold-start issue
Bied et al. [11] Public Job Recommendation	MUSE.2 (Two-Tier Multi-Head Neural Re-ranker)	XGBoost (RecSys challenge-winning baseline)	MUSE.2 Recall@10 = 30.1%, Recall@20 = 40.2% vs XGBoost Recall@10 = 26.83%, Recall@20 = 35.59%	User cold-start test on RecSys 2017 data	MUSE.2 Recall@100: 24.9% (warm) to 24.0% (cold); DropoutNet: 41.2% to 23.1%. Standalone modules on Pôle emploi data: geographic ~15%, general text ~34%

Study Reference & Decision Task	Proposed Architecture	Best Competitive Baseline	Baseline vs Proposed Metrics	Empirical Stress Test / Failure Mode Evaluated	Observed Performance Degradation / Fairness Penalty
Senger et al. [2] Career Trajectory	MLP / Linear with fine-tuned MPNet embeddings	Meta-Llama-3-8B-Instruct (LLM approach)	MLP MRR = 43.58 vs LLM MRR = 34.16 on K+occ	Titles-only inputs; synthetic data errors; LLM fine-tuning	Titles-only reduced Linear MRR on K+ESCO from 47.75 to 38.87; one-epoch fine-tuning of Llama-3 on DECORTE: 18.2 to 13.96
Le et al. [5] Career Intelligence	CareerPathKG (Graph-guided local inference)	GPT-5 / Gemini-2.5-Pro (Commercial frontier LLMs)	CareerPathKG RMSE = 8.9, Recall@10 = 78.3% vs GPT-5 RMSE = 12.4, Recall@10 = 68.9%	Inference latency & semantic overlap error analysis	Commercial LLM latency reached 13.9s (GPT-5) and 22.4s (Gemini-2.5-Pro) per CV-JD matching query (vs 0.24s for KG); KG exhibited false-positive errors from semantic over-matching (e.g. data vs CI/CD pipelines)
Sühr et al. [14] Gig Platform Hiring	FairDet-Greedy (Post-hoc fair reranking)	RabbitRanking (Standard platform relevance sort)	Female share of first selections, D1 Moving: FairDet-Greedy 12.3% vs RabbitRanking 4.8%	Multi-choice human screener degradation (evaluating employer selections at Choices 2, 3, and 4)	Gap between rankings decreased as selections increased (Moving, all sets: 17.4 to 5.2 percentage points); authors find less impact on the overall fraction of female candidates selected

Study Reference & Decision Task	Proposed Architecture	Best Competitive Baseline	Baseline vs Proposed Metrics	Empirical Stress Test / Failure Mode Evaluated	Observed Performance Degradation / Fairness Penalty
Peng et al. [16] Human-AI Screening	Hybrid Human + DNN (Deep representation model)	Hybrid Human + Bag-of-Words (Interpretable sparse linear model)	Human + DNN teams unbiased; Human + BOW teams reflected human and model biases (e.g., surgeon task biased toward men)	Conformity to erroneous algorithmic recommendations (evaluating screener choices when model is incorrect)	Matched incorrect model decisions in 70.4-80.9% of cases with BOW vs 50.4-66.3% with DNN
Booth et al. [19] Video Interviewing	Verbal-Only Model (Text transcripts from audio)	Multimodal Combined Model (Text + Acoustic + Computer Vision)	Verbal alone Spearman ρ = 0.45, AIR = 0.87 vs Multimodal ρ = 0.46, AIR = 0.76; Visual AIR = 0.52	Modality ablation and demographic leak test (AUROC of predicting candidate gender from features)	Adding video/audio offered a negligible accuracy gain (ρ = 0.46 vs 0.45); gender AUROC rose from 0.73 (verbal) to 0.98 (combined); visual AIR 0.52 and combined AIR 0.76 were below the 0.80 benchmark

4 Discussion

4.1 A common evaluation ladder for unlike models

The empirical findings compiled in Table 4 and Table S1 demonstrate why conventional offline metric leaderboards (e.g., comparing AUC, MRR, or F1 scores) are scientifically uninformative when divorced from operational context. A model reporting an AUC of 0.95 on recruiter decisions [9] is not 'superior' to a model reporting an AUC of 0.93 on balanced job-offer outcomes [8], nor does a skill recommender with 98.7% overall accuracy [3] represent a practical solution when its F1-score is below 0.50.

To structure future research and deployment audits, we formalize a four-level evaluation ladder for career and talent AI, illustrated in Fig. 9:

- Level 1: Representation Validity. Assesses whether extracted tokens, vector embeddings, and taxonomy mappings accurately reflect the source phenomenon without systematic omission or artifactual distortion. Exemplified by skill extraction benchmarks [3, 12] and psychometric classifications [6], Level 1 confirms that computational representations are internally consistent. However, Level 1 evaluations cannot establish whether an inferred skill or personality trait predicts genuine vocational competence.

- Level 2: Offline Decision-Stage Validation. Evaluates predictive performance on held-out candidate records, vacancies, or application cycles under realistic operational constraints. As demonstrated by Qin et al. [8], Jiang et al. [9], and Bied et al. [11], Level 2 evaluations establish predictive utility against historical operational labels. However, Level 2 validity is fragile: models that succeed under random cross-validation can degrade under temporal shifts, sparse career histories, or user cold starts.

- Level 3: Human-System Interaction Assessment. Investigates how algorithmic outputs, explainability displays, and fairness interventions alter human cognitive understanding, trust calibration, and joint decision outcomes. Exemplified by Sühr et al. [14], Wang et al. [15], and Peng et al. [16], Level 3 evaluations examine the socio-technical interface. The evidence indicates that mathematical fairness at Level 2 does not guarantee equitable human choices at Level 3, and screeners may conform more to more interpretable models, with bias outcomes differing by model.

- Level 4: Consequential Impact Evaluation. Measures downstream, prospective real-world outcomes over longitudinal horizons, including interview conversion rates, job placement durability, employee retention, wage trajectories, job satisfaction, and disparate career impacts across protected demographic groups.

Strikingly, none of the 19 primary close-read studies conducted a rigorous Level 4 prospective evaluation. Whether any of the remaining 162 candidate studies does so has not been verified at full-text level and is a priority for future work. Within the appraised evidence, the literature operates almost entirely within Levels 1 and 2, with rare excursions into Level 3. Advancing the field requires bridging this gap between offline proxy optimization and consequential real-world impact.

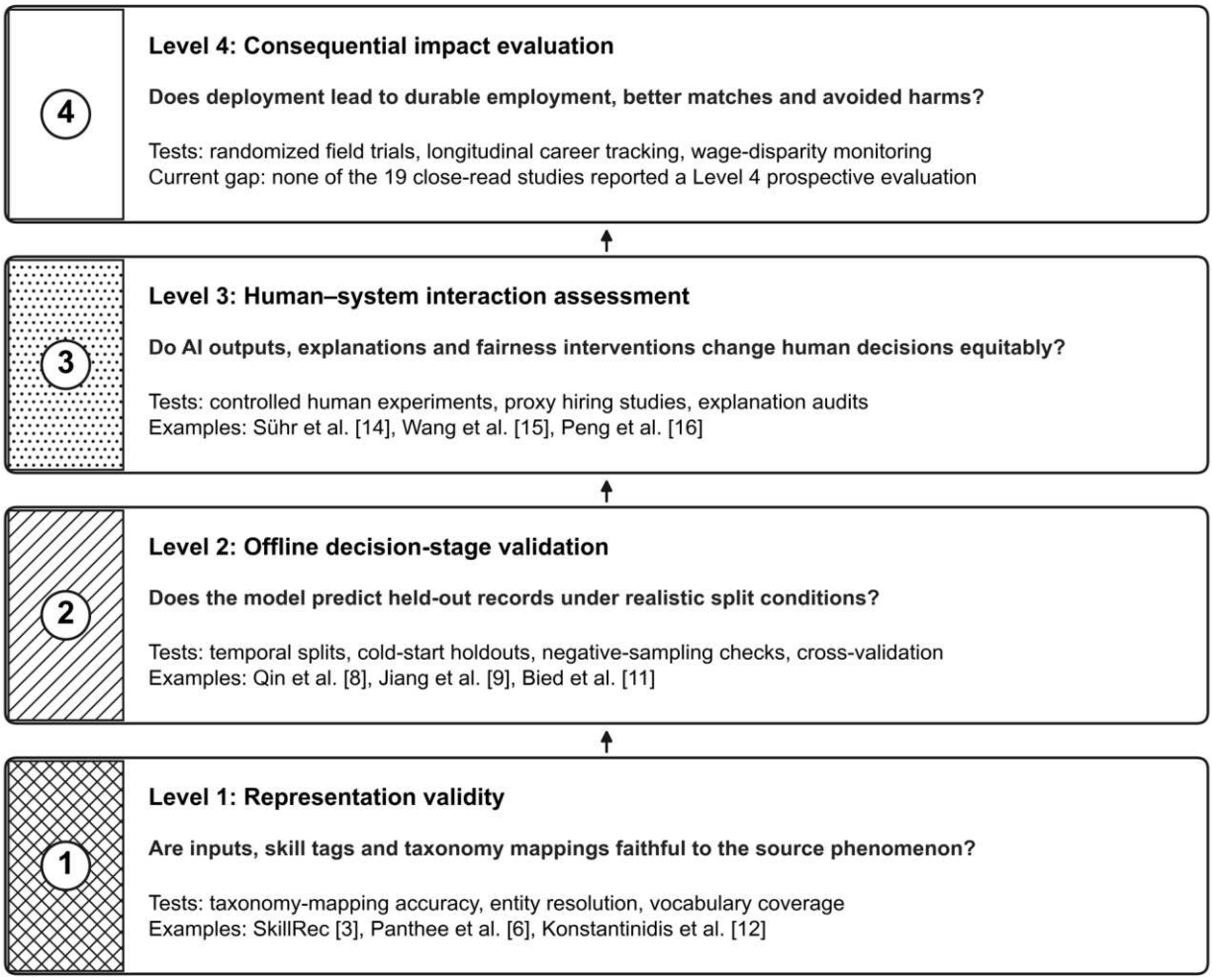

Fig. 9 Four-level evaluation ladder for career and talent AI systems. Hierarchical evaluation progression from Level 1 (Representation Validity) through Level 2 (Offline Decision-Stage Validation) and Level 3 (Human-System Interaction Assessment) to Level 4 (Consequential Impact Evaluation). Evidence at lower levels cannot establish validity at subsequent levels.

4.2 Practical stress-testing and methodological recommendations

Based on the failure modes and empirical vulnerabilities identified across the reviewed studies, we propose five concrete methodological recommendations for researchers, enterprise developers, and regulatory auditing bodies:

1. Mandatory Cold-Start and Temporal Partitioning: Research reporting person-job fit or trajectory prediction must evaluate models under explicit user/job cold-start splits and strict chronological holdouts, rather than relying exclusively on random cross-validation. Developers must report performance disaggregated by candidate career tenure, specifically reporting metrics for job seekers with few recorded work experiences, as Gong et al. did for those with fewer than three [10].

2. Construct Validation of Taxonomies and Psychometrics: Studies utilizing occupational taxonomies (e.g., ESCO, O*NET) must measure and report mapping error rates, out-of-

vocabulary frequencies, and performance across minority or emerging job sectors, rather than treating taxonomy assignment as self-evident ground truth [2, 12]. Psychometric profiling systems must provide empirical evidence linking questionnaire classifications to longitudinal vocational outcomes [6].

3. Point-of-Impact Fairness and Behavioral Audits: Algorithmic fairness must be evaluated at the point of ultimate decision impact rather than stopping at model-level diagnostic parity. Audits of candidate ranking systems must test multi-item selection dynamics (evaluating choices across ranks 1 through k) in realistic human screening environments to assess whether fair rankings withstand human cognitive shortcuts [14, 18].

4. Prudent Modality Selection and Bias Auditing: In automated video interviewing and candidate assessment, non-verbal sensing (facial video analysis, acoustic emotion recognition) should be avoided unless compelling, peer-reviewed evidence demonstrates significant incremental validity over text transcripts. In light of Booth et al. [19], multimodal video scoring increased demographic bias, with visual and combined models below the four-fifths benchmark, for only a negligible predictive gain in that study.

5. De-biasing Reference Distribution Alignment: Audits of commercial screening algorithms should assess whether the groups used for adverse impact testing represent actual job seekers, since otherwise bias assessments may be flawed [17].

4.3 Scope, missing evidence and reproducibility

This structured methodological review is subject to three specific limitations that should be explicitly considered when interpreting its findings:

1. Search Scope and Retrieval Attrition: Literature retrieval was conducted through a single search platform, Elicit, using two targeted semantic search queries returning 1,000 results in total. 409 of the 715 sought reports (57.2%) were recorded as full text not available to the screening tool, with no reason given for individual cases. Descriptive metadata comparison shows a similar recency profile for unretrieved reports with a known year (250 of 401; 62.3%) and assessed reports (62.4%) published in 2025-2026, but higher rates of missing venue (10.5% vs 1.0%) and DOI metadata (15.4% vs 0.0%) among the unretrieved. These differences are consistent with electronic indexing latency, but this descriptive comparison cannot rule out systematic differences, and the unretrieved reports represent an unmeasured fraction of the literature. Prevalence figures for tasks, architectures and years (Figs. 2 to 4) should therefore be read as descriptive of the retrievable corpus, not of the field as a whole.

2. Candidate Extraction Set Status: The 181 candidate extraction studies were identified through screening using Elicit and coded by the authors from Elicit-generated extraction text rather than from the full papers, with independent double coding on a random sample of 30 studies (Section 2.3). They have not undergone independent, dual-human re-adjudication of publication venue authenticity or claim-level verification. Consequently, the candidate set represents a traceable candidate evidence pool rather than a fully verified systematic census.

3. Purposive Primary Set Selection: The 19 primary close-read studies were purposively selected to represent maximum methodological diversity across tasks, architectures, datasets, and trustworthiness dimensions. While these studies provide rich, inspectable empirical evidence, they do not constitute a statistically random sample of the 181 candidate studies. Therefore, quantitative proportions calculated from the 19 studies cannot be generalized to the broader literature. In particular, 15 of the 19 studies (79%) were published in 2020-2023, against 62.4% of the candidate set in 2025-2026, and only two primary studies evaluate LLM-based systems [2, 5]; conclusions about transformer and LLM approaches are therefore preliminary and should be revisited with a close-read set weighted toward recent work.

5 Conclusions

Artificial intelligence in career guidance and talent matching cannot be evaluated as a monolithic technology. By deconstructing the full decision pipeline across 181 candidate extraction studies and 19 primary close-read investigations, this review establishes that model efficacy is profoundly shaped by the operational choice of ground-truth labels, validation partitioning designs, and downstream human interaction dynamics.

Our findings demonstrate that high offline accuracy scores frequently mask critical operational vulnerabilities: performance degradation for job seekers with few recorded work experiences, weak graph baselines that ignore content, sensitivity to synthetic data quality and small-sample fine-tuning in language models, and the attenuation of algorithmic fairness interventions under human decision-making. Furthermore, in the one appraised multimodal study, video interview assessment increased demographic bias with only a negligible increment in accuracy over text alone.

Future progress requires transitioning away from static offline metric leaderboards toward rigorous, multi-level evaluation frameworks. By adopting the four-level evaluation ladder, linking representation validity, offline decision-stage stress-testing, human-system interaction assessment, and consequential real-world impact, the field can advance toward vocational AI systems that are technically robust, methodologically valid, and socially equitable.

Supplementary Information

ESM_1 (Excel workbook): verbatim search queries, screening criteria, extraction instructions, title and abstract screening records for 980 records, full-text screening records for 715 reports, and the selection reconciliation.

ESM_2 (Excel workbook): extraction records for the 181 candidate studies, final author labels, the coding codebook, the double-coded sample of 30 studies with its resolved labels, the master synthesis of the 19 primary studies (Table S1), and the traceability crosswalk (Table S2).

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests: The authors declare that they have no competing interests.

Author contributions: A.N. contributed to defining the research methodology and structuring the literature review; wrote the screening criteria and codebook; analyzed and compared research papers, examining the different ML and AI methodologies, evaluating their strengths and limitations, and identifying research gaps; and coded all 181 candidate studies. H.S. contributed to the literature review by investigating the selected research papers, examining their proposed approaches, comparing methodologies, and analyzing their findings, limitations, and potential improvements; and independently coded the random sample of 30 studies. M.F. supervised the entire research process, provided academic guidance on research methodology, literature review structure and research direction, defined the key aspects to investigate, and helped refine the scope and focus of the review. All authors read and approved the final manuscript.

Ethics approval and consent to participate: Not applicable. No new human or animal participant data were collected for this secondary literature review.

Consent for publication: Not applicable.

Use of Artificial Intelligence: Elicit (Elicit Research) was used to support literature retrieval, title and abstract screening, full-text screening, and data extraction. The authors take full responsibility for the content of this manuscript, including screening decisions, coding, data analysis, interpretation of results, and scientific conclusions. Beyond the use of Elicit described above, any other use of artificial intelligence tools was limited to language refinement and editorial assistance and did not contribute to the study design, independent scientific judgment, interpretation of results, or scientific conclusions.

Data availability: The verbatim search queries, screening criteria, screening records, extraction instructions, extraction data, codebook and coding labels supporting this review are provided in the Supplementary Information (ESM_1 and ESM_2). The full Elicit report is available from the corresponding author on request.

Code availability: The Python scripts used for statistics are available from the corresponding author on request.

Clinical trial number: not applicable.

References

[1] Ghosh A, Woolf B, Zilberstein S, Lan AS (2020) Skill-based career path modeling and recommendation. In: Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), pp 1156-1165. https://doi.org/10.1109/BigData50022.2020.9377992

[2] Senger E, Campbell Y, van der Goot R, Plank B (2025) Toward more realistic career path prediction: evaluation and methods. Frontiers in Big Data 8:1564521. https://doi.org/10.3389/fdata.2025.1564521

[3] Ong XQ, Lim KH (2023) SkillRec: a data-driven approach to job skill recommendation for career insights. In: Proceedings of the 2023 15th International Conference on Computer and Automation Engineering (ICCAE), pp 40-44. https://doi.org/10.1109/ICCAE56788.2023.10111438

[4] Guleria P, Sood M (2022) Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. Education and Information Technologies 28(1):1081-1116. https://doi.org/10.1007/s10639-022-11221-2

[5] Le N-Q, Hoang DD, Vuong T-H-Y, Tran MV (2026) CareerPathKG: knowledge graph integrated framework for career intelligence. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Industry Track, Volume 5: Industry Track, pp 813-822. https://doi.org/10.18653/v1/2026.eacl-industry.60

[6] Panthee S, Rajkarnikar S, Begum R (2023) Career guidance system using machine learning. Journal of Advanced College of Engineering and Management 8(2):113-119. https://doi.org/10.3126/jacem.v8i2.55947

[7] Gugnani A, Misra H (2020) Implicit skills extraction using document embedding and its use in job recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence 34(08):13286-13293. https://doi.org/10.1609/aaai.v34i08.7038

[8] Qin C, Zhu H, Xu T, Zhu C, Ma C, Chen E, Xiong H (2020) An enhanced neural network approach to person-job fit in talent recruitment. ACM Transactions on Information Systems 38(2):1-33, Article 15. https://doi.org/10.1145/3376927

[9] Jiang J, Ye S, Wang W, Xu J, Luo X (2020) Learning effective representations for person-job fit by feature fusion. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), pp 2549-2556. https://doi.org/10.1145/3340531.3412717

[10] Gong Z, Song Y, Zhang T, Wen J-R, Zhao D, Yan R (2024) Your career path matters in person-job fit. In: Proceedings of the AAAI Conference on Artificial Intelligence 38(8):8427-8435. https://doi.org/10.1609/aaai.v38i8.28685

[11] Bied G, Nathan S, Perennes E, Hoffmann M, Caillou P, Crépon B, Gaillac C, Sebag M (2023) Toward job recommendation for all. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI '23), pp 5906-5914. https://doi.org/10.24963/ijcai.2023/655

[12] Konstantinidis I, Maragoudakis M, Magnisalis I, Berberidis C, Peristeras V (2022) Knowledge-driven unsupervised skills extraction for graph-based talent matching. In:

Proceedings of the 12th Hellenic Conference on Artificial Intelligence (SETN '22), pp 1-7. https://doi.org/10.1145/3549737.3549769

[13] Delecraz S, Eltarr L, Becuwe M, Bouxin H, Boutin N, Oullier O (2022) Making recruitment more inclusive: unfairness monitoring with a job matching machine-learning algorithm. In: Proceedings of the 2022 IEEE/ACM International Workshop on Equitable Data & Technology (FairWare '22), pp 34-41. https://doi.org/10.1145/3524491.3527309

[14] Sühr T, Hilgard S, Lakkaraju H (2021) Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), pp 989-999. https://doi.org/10.1145/3461702.3462602

[15] Wang C, Wang K, Bian A, Islam R, Keya K, Foulds JR, Pan S (2023) When biased humans meet debiased AI: a case study in college major recommendation. ACM Transactions on Interactive Intelligent Systems 13(3):1-28, Article 17. https://doi.org/10.1145/3611313

[16] Peng A, Nushi B, Kıcıman E, Inkpen K, Kamar E (2022) Investigations of performance and bias in human-AI teamwork in hiring. In: Proceedings of the AAAI Conference on Artificial Intelligence 36(11):12089-12097. https://doi.org/10.1609/aaai.v36i11.21468

[17] Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J, Trindel K, Polli F (2021) Building and auditing fair algorithms: a case study in candidate screening. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp 666-677. https://doi.org/10.1145/3442188.3445928

[18] Behzad T, Devic S, Sharan V, Korolova A, Kempe D (2026) An external fairness evaluation of LinkedIn Talent Search. In: Proceedings of the AAAI Conference on Artificial Intelligence 40(45):38224-38232. https://doi.org/10.1609/aaai.v40i45.41161

[19] Booth BM, Hickman L, Subburaj SK, Tay L, Woo SE, D'Mello SK (2021) Bias and fairness in multimodal machine learning: a case study of automated video interviews. In: Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), pp 268-277. https://doi.org/10.1145/3462244.3479897

[20] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic Reviews 10:89. https://doi.org/10.1186/s13643-021-01626-4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.xlsx
- ESM2.xlsx