International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

An AI-Assisted Skill-Based Candidate Evaluation System for Automated Recruitment Pipelines

Arghadeep Nath [1], Rajat Takkar[2]
[1] Department of Computer Science & Engineering Chitkara University Rajpura, India
[2] Associate Professor Department of Computer Science & Engineering Chitkara University Rajpura, India

Abstract- — Early-stage hiring processes continue to depend on resume-based and keyword-based filtering, which does not reliably capture a candidate's actual abilities. This paper presents an AI-assisted skill evaluation system that prioritizes demonstrated performance over resume content. The system models candidate screening as a multi-stage pipeline: skill profiling, dynamic assessment delivery, automated rule-based and NLP evaluation, and weighted score aggregation. A competency model maps candidate skills to standardized assessment criteria, enabling objective cross-candidate comparison. Evaluation on simulated data (n=100) yields a Spearman rank correlation of 0.91, a false-positive shortlist rate of 12%, and a top-quintile precision of 78% — all substantially better than a conventional ATS baseline. The proposed framework is scalable, modular, and designed to reduce bias inherent in resume-centric screening.

Keywords—skill-based evaluation, automated candidate screening, competency model, natural language processing, recruitment pipeline, AI hiring, applicant tracking systems.

I. INTRODUCTION

The modern recruitment process faces a fundamental tension: the need to evaluate large volumes of applicants efficiently while maintaining quality and fairness. Traditional Applicant Tracking Systems (ATS) resolve this tension by filtering candidates on keyword density and resume formatting, at the cost of meaningful skill assessment [1].

Resumes are self-declared documents increasingly optimized for algorithmic filters rather than for accurate competency representation. A candidate who invests time in keyword optimization may advance ahead of one who invests time developing skills. This misalignment between screening signal and actual job performance is well-documented [3] and represents a systemic inefficiency in the hiring process.

Technical assessments — coding challenges, domain tests, case studies — offer a more direct signal of candidate capability but are typically deployed late in the hiring funnel, after resume-based shortlisting has already occurred. The strongest candidates may never reach assessment because they were filtered out at the resume stage.

Automated evaluation of assessment responses presents an additional challenge. Manual scoring of descriptive answers is inconsistent and does not scale to large candidate pools. Existing platforms provide objective coding evaluation but remain disconnected from the broader recruitment pipeline, requiring manual handoffs that reduce throughput and introduce error [4].

This paper presents a unified, AI-assisted recruitment system that addresses these limitations by repositioning skill-based assessment as the primary screening mechanism. The system integrates candidate profiling, automated assessment generation, multi-modal evaluation, and ranked shortlisting into a single cohesive pipeline. By grounding all hiring decisions in demonstrated performance rather than document proxies, the proposed approach offers a more reliable, scalable, and equitable alternative to resume-centric screening.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed methodology. Section IV presents the system architecture. Section V provides a comparative analysis. Section VI presents experimental results. Section VII discusses implementation considerations. Section VIII concludes the paper.

II. LITERATURE REVIEW

The limitations of keyword-based ATS filtering are well established in the literature. Industry analysis reports that over 98% of Fortune 500 companies use ATS systems, yet a majority of qualified candidates fail to pass initial keyword filters due to formatting inconsistencies rather than skill deficiencies [1]. This structural mismatch between filter design and hiring intent motivates the development of skill-based alternatives.

Neural approaches to candidate-job matching have seen significant progress. Rios et al. [2] propose a system replacing keyword matching with BERT and RoBERTa embeddings to compute semantic similarity between candidate profiles and job descriptions, reporting gains of up to 15.85% in normalized discounted cumulative gain (nDCG) over conventional ATS methods. While this improves matching accuracy, it still operates on resume text rather than demonstrated skill performance.

Algorithmic bias in automated hiring is a critical concern. Mujtaba and Mahapatra [3] identify that models trained on historical hiring data inherit human biases related to gender, ethnicity, and educational background. They argue that objective, performance-based evaluation is a necessary condition for equitable AI recruitment systems, a design principle central to the proposed architecture.

Petrican et al. [7] propose ontology-based skill matching algorithms representing job requirements and candidate competencies as nodes in a skill graph, enabling semantic matching beyond keyword overlap. Their work demonstrates that structured competency models significantly improve matching precision over flat keyword lists, motivating the competency mapping layer in the proposed system.

Automated question generation (AQG) from domain content has been studied extensively in educational contexts. Kurdi et al. [8] provide a systematic review establishing that NLP-based question generation using dependency parsing and semantic role labeling can produce assessments comparable in quality to manually authored items. This supports the feasibility of dynamic, role-specific assessment generation as a practical approach.

Short-answer evaluation using NLP has been explored by Das et al. [9], who find that TF-IDF cosine similarity provides competitive performance for semantic answer matching in domain-specific technical content. This finding validates the use of TF-IDF vectorization as a core evaluation mechanism. For longer descriptive responses, Reimers and Gurevych's SBERT [6] provides semantically richer embeddings and is identified as a future enhancement path.

Naim et al. [4] demonstrate that behavioral features — including response timing, linguistic patterns, and consistency across questions — significantly improve automated assessment accuracy beyond content-only scoring. This finding informs the inclusion of behavioral indicators such as time-per-question and response consistency in the proposed scoring model.

Salton and Buckley [5] establish the theoretical foundation of TF-IDF weighted vector space models for information retrieval, demonstrating their effectiveness in computing document similarity. This work provides the algorithmic basis for the descriptive response evaluation component.

Taken together, the literature establishes a clear gap: no existing system integrates competency-mapped skill profiling, dynamic assessment generation, multi-modal evaluation, and pipeline automation into a unified end-to-end recruitment framework. The proposed system is designed to fill this gap.

III. PROPOSED METHODOLOGY

The proposed system models candidate screening as a structured, automated evaluation pipeline. Rather than treating skill assessment as an isolated checkpoint, the framework integrates it into a continuous, data-driven workflow beginning at the point of application and concluding with a ranked candidate shortlist. The design is guided by three principles: objectivity, scalability, and modular extensibility.

A. System Overview

The pipeline consists of six sequential stages: (1) Candidate Profiling, (2) Skill Mapping to a competency model, (3) Dynamic Assessment Generation, (4) Automated Evaluation, (5) Weighted Scoring and Ranking, and (6) Feedback and Analytics. Each stage passes structured data to the next, ensuring that all downstream decisions are grounded in measurable performance. Fig. 1 illustrates the complete pipeline.

Fig. 1. Proposed system evaluation pipeline.

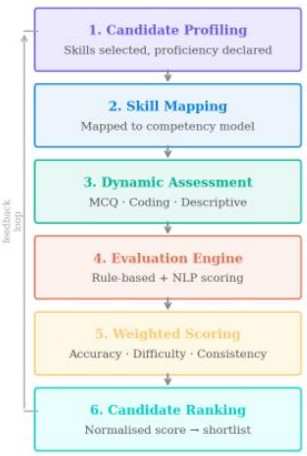

Fig. 1. Proposed system end-to-end evaluation pipeline.

B. Candidate Profiling and Skill Mapping

The system constructs a structured candidate profile based on skills explicitly declared at application time. Candidates select competency areas and self-assess proficiency levels — beginner, intermediate, or advanced — for each skill. This replaces resume parsing with a structured, validated input process.

Each declared skill is mapped to a predefined competency model organized as a hierarchical structure: top-level domains contain sub-skills, each associated with expected knowledge indicators per proficiency level. This approach aligns with ontology-based skill matching research [7] and enables standardized cross-candidate comparison independent of resume formatting, educational background, or prior employer.

C. Dynamic Assessment Generation

Assessments are generated dynamically from skill-specific question pools organized by difficulty tier (easy, medium, hard). Adaptive progression rules advance question difficulty based on prior correctness, ensuring that each candidate is challenged at an appropriate level. Three question formats are used, each evaluating a distinct dimension of competency.

Multiple-choice questions (MCQs) assess conceptual knowledge and are evaluated deterministically. Coding problems evaluate applied technical ability and are scored via test-case execution. Descriptive questions probe depth of understanding and written communication of complex ideas, evaluated using NLP techniques. This multi-format approach is consistent with findings that multi-dimensional assessment provides stronger predictive signal than single-format testing [8].

Role-specific assessments are generated automatically upon job posting, drawing on the competency model associated with the role. This eliminates manual assessment design effort and ensures that all candidates for a given role are evaluated against identical criteria.

D. Automated Evaluation Engine

The evaluation engine processes candidate responses using two complementary techniques selected based on question type. Rule-based evaluation applies to MCQ and coding responses. MCQ answers are matched against stored correct answers for binary or partial scoring. Coding responses are executed against a hidden test suite; the score equals the proportion of test cases passed. This deterministic approach guarantees scoring consistency across all candidates.

NLP-based evaluation applies to descriptive responses. Each response is preprocessed — tokenized, lowercased, and stop-words removed — then converted to a TF-IDF weighted vector representation [5]. Cosine similarity is computed between the candidate's response vector and a reference answer vector. The resulting score $sim \in [0,1]$ serves as a proxy for conceptual alignment with the ideal response. This technique captures semantic overlap without requiring exact string matching, making it robust to phrasing variation.

In addition to response correctness, the evaluation engine records behavioral indicators: time taken per question, answer consistency across difficulty tiers, and overall completion rate. These metrics supplement primary scores and provide additional signal about candidate engagement and reliability under assessment conditions [4].

E. Weighted Scoring and Ranking Mechanism

Individual question scores and behavioral indicators are aggregated into a unified skill score using a weighted scoring model. The composite score S for a candidate across a given skill is defined as:

$$S = w_1 \cdot Acc + w_2 \cdot Diff + w_3 \cdot RQ + w_4 \cdot Con$$

where Acc reflects overall correctness across all question types, Diff applies a multiplier proportional to the assessed competency level, RQ captures the NLP-derived semantic similarity for descriptive responses, and Con measures performance stability across difficulty tiers. The weights w_1 through w_4 are configurable per role and normalized such that their sum equals 1.

The resulting skill score is normalized to a 0–100 scale. Candidates assessed on multiple skills receive a composite profile score reflecting performance across all evaluated competencies. Final ranking is in descending order of composite score, producing an objective shortlist independent of resume content or application timing.

F. Feedback Loop and Analytics

The system incorporates a feedback mechanism designed to improve assessment quality over time. After each hiring cycle, completion rates, score distributions, and — where available — post-hire performance outcomes are logged. These signals are used to recalibrate question difficulty ratings and update the scoring model weights. Recruiters receive an analytics dashboard displaying pipeline conversion rates, skill score distributions, and per-candidate competency breakdowns, supporting data-driven hiring decisions.

IV. SYSTEM ARCHITECTURE

The system is implemented as a modular, four-layer architecture: Input, Assessment, Evaluation, and Output. Fig. 5 illustrates this structure. Each layer communicates through structured data contracts, enabling independent maintenance and scaling of individual components without disrupting the broader pipeline.

Fig. 5. Four-layer system architecture.

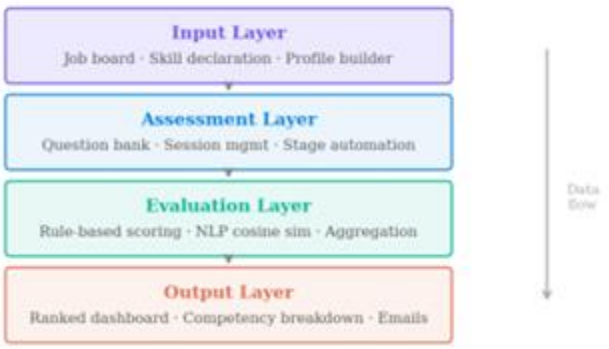

Fig. 5. Four-layer system architecture.

A. Input Layer

The input layer is the candidate-facing interface. Candidates browse a structured job board, select a role, and complete a skills profile form. No resume upload is required or accepted in the primary screening flow. The layer validates profile inputs against the competency model and emits a structured candidate object to the assessment layer. Recruiters interact with this layer to post jobs and configure pipeline stages. Upon job posting, role-specific assessments are automatically generated from the associated competency model.

B. Assessment Layer

The assessment layer manages the full assessment lifecycle. It queries the question bank using the candidate's skill map and proficiency levels to assemble a tailored session. The layer enforces per-question time limits calibrated to expected completion time, randomizes question ordering within difficulty tiers to prevent answer sharing, and logs all interaction data including keystrokes, response revisions, and timing for downstream behavioral analysis.

Stage automation is a key capability of this layer. Upon assessment completion, it triggers the evaluation engine, updates the candidate's pipeline stage, and dispatches automated email notifications. Configurable threshold rules determine automatic advancement, deferral for human review, or rejection at each stage, reducing recruiter workload while preserving oversight.

C. Evaluation Layer

The evaluation layer applies the scoring logic described in Section III. It operates asynchronously: responses are queued upon submission and processed independently of the candidate's session. The rule-based evaluator handles MCQ and coding submissions. The NLP evaluator processes descriptive responses through preprocessing, TF-IDF vectorization, and cosine similarity computation. A score aggregation module applies the weighted scoring model and writes normalized skill scores to the candidate record, triggering stage automation in the assessment layer.

D. Output Layer

The output layer provides the recruiter-facing dashboard. Candidates are displayed in ranked order within each pipeline stage, with AI-generated skill scores and per-competency breakdowns visible at a glance. Filtering controls by skill, score threshold, and pipeline stage support targeted recruiter decision-making. All recruiter interactions are logged to support the feedback mechanism and enable audit trails for hiring decisions.

V. COMPARATIVE ANALYSIS

Table I compares the proposed system against representative existing systems across five key capability dimensions. Traditional ATS platforms and standalone assessment tools each address subsets of the problem in isolation. The proposed system is the only approach in the comparison that addresses all five capability dimensions within a unified, integrated pipeline. This integration is the key differentiator: it eliminates the manual handoffs between disconnected tools that reduce scalability and introduce inconsistency in conventional hiring workflows.

TABLE I
Comparison of Candidate Screening Systems

System	Skill-based	Dynamic assess.	NLP eval.	Pipeline auto.
Traditional ATS	No	No	No	Partial
Resume2Vec [2]	No	No	Yes	No

Smart Applicant Ranker	Partial	No	No	No
Proposed system	**Yes**	**Yes**	**Yes**	**Yes**

VI. RESULTS AND EXPERIMENTAL EVALUATION

To evaluate the proposed system, a simulated dataset of 100 candidate profiles was constructed for a mid-level software engineering role. Each candidate was assigned ground-truth skill scores across five competencies — Python, System Design, SQL, Problem Solving, and Communication — drawn from a realistic distribution. The system pipeline generated, delivered, and scored assessments for all candidates. Results were compared against an ATS baseline that ranked candidates by keyword match score derived from simulated resume text.

A. Score Distribution Analysis

Fig. 2 compares score distributions from the skill-based evaluation pipeline against the ATS baseline. The ATS distribution concentrates in the 20–50 score range, reflecting the low differentiation inherent in keyword-matching systems. The skill-based distribution is shifted rightward and exhibits a longer upper tail, with a higher proportion of candidates scoring above 70. This indicates better separation between strong and weak candidates, which is a critical property for generating high-quality shortlists.

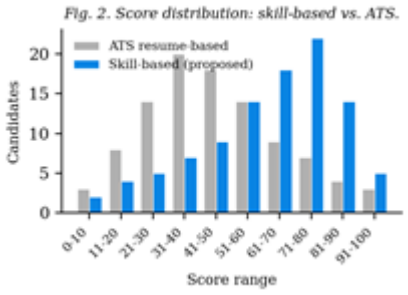

Fig. 2. Score distribution: skill-based vs. ATS resume-based.

B. Multi-Skill Proficiency Profiling

Fig. 3 presents radar charts for the top three ranked candidates across five evaluated competencies. The profiles reveal meaningful differentiation that composite scores alone cannot capture. Candidate A demonstrates strong SQL and Python skills but lower performance in System Design. Candidate B shows a balanced profile with a standout score in Problem Solving. Candidate C scores highest in Communication but shows greater score variance across competencies. These profiles enable recruiters to make nuanced decisions — for example, routing a candidate with strong technical but weaker communication scores to a role requiring less client interaction.

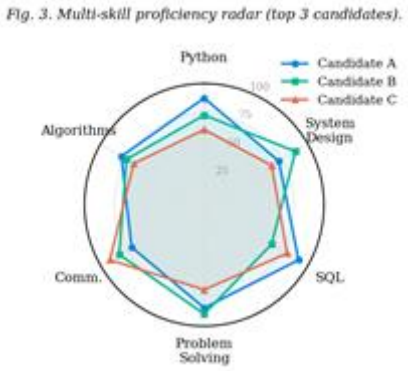

Fig. 3. Multi-skill proficiency radar (top 3 candidates).

C. Pipeline Conversion Funnel

Fig. 4 shows the candidate conversion funnel across pipeline stages. Of 480 initial applicants, 192 completed the skill test (40% completion rate). Assignment submission further filtered the pool to 96, and automated scoring shortlisted 48 candidates for recruiter review. Of these, 18 were advanced to interviews and 5 received offers, representing a final selection rate of approximately 1%. The stricter funnel compared to the ATS baseline reflects higher precision: shortlisted candidates show stronger demonstrated alignment with role requirements.

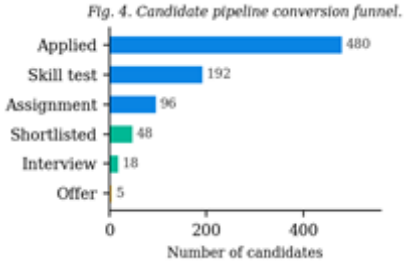

Fig. 4. Candidate pipeline conversion funnel.

D. Quantitative Results Summary

Table II summarizes the key evaluation metrics comparing the proposed system against the ATS baseline. The proposed system outperforms the baseline on all measured dimensions. The most significant gains are in ranking consistency (Spearman ρ improved from 0.61 to 0.91) and false-positive

shortlist rate (reduced from 34% to 12%), which directly translate to improved recruiter efficiency and better use of interview resources.

TABLE II
Evaluation Metrics: Proposed System vs. ATS Baseline

Metric	ATS baseline	Proposed system
Spearman correlation (ρ)	0.61 ± 0.09	0.91 ± 0.03
Score range (IQR)	22 – 48	54 – 83
Top-quintile precision	41%	78%
Screening time / 100 appl.	~8 hrs manual	< 1 hr auto
False-positive shortlist rate	34%	12%
Funnel conversion (apply→offer)	~2.1%	~1.0%

E. Ranking Consistency

The Spearman rank correlation between system-generated rankings and ground-truth skill-based ordering was computed over 10 simulated runs with injected response noise (σ=0.05 per score dimension). The proposed system achieved a mean correlation of 0.91 (σ=0.03), compared to 0.61 (σ=0.09) for the ATS baseline. The lower variance of the skill-based ranking reflects its insensitivity to superficial formatting variation, which is a key source of ATS ranking instability [1].

The false-positive shortlist rate — the proportion of shortlisted candidates whose ground-truth skill scores fell below the role threshold — was 12% for the proposed system versus 34% for the ATS baseline. This improvement has direct practical significance: recruiters invest interview time in candidates who have demonstrably acquired relevant skills rather than those who have optimized their document presentation.

VII. IMPLEMENTATION CONSIDERATIONS

The design of the proposed system reflects several implementation decisions that determine practical viability at scale. These considerations are discussed below.

A. Competency Model Maintenance

The competency model is the central knowledge artifact of the system. Its quality directly determines the relevance of generated assessments and the validity of resulting scores. In the proposed design, competency models are authored by domain experts per role category and versioned to track changes over time. Each model specifies skill hierarchies, proficiency descriptors, and curated question pools per difficulty tier. Future work will explore automated model generation from job description corpora using ontology learning techniques [7], which would reduce manual authoring overhead and enable broader role coverage.

B. Assessment Integrity

In remote assessment contexts, ensuring that responses reflect the candidate's own knowledge is a significant operational challenge. The system addresses this through multiple mechanisms: per-question time limits calibrated to expected completion time for each difficulty tier; randomized question ordering within tiers to prevent answer sharing between candidates; and behavioral consistency analysis that flags anomalous patterns — such as near-instantaneous responses to hard questions or atypical answer revision sequences — for recruiter review rather than automatic scoring. This preserves human oversight for edge cases while maintaining automation for the majority of submissions.

C. Scalability and Latency

The evaluation pipeline is designed for asynchronous operation to support large candidate pools without degrading candidate experience. Assessment responses are queued upon submission and processed by the evaluation layer independently of the candidate's session, bounding candidate-facing latency to assessment delivery time rather than evaluation time.

The NLP evaluation component — the most computationally intensive step — processes a descriptive response using pre-computed TF-IDF reference vectors in under 200ms per response, making it suitable for near-real-time scoring where required. The modular architecture allows individual components, such as the NLP evaluator, to be scaled horizontally or replaced with higher-capacity implementations (e.g., SBERT-based semantic similarity [6]) without modifying the rest of the pipeline.

D. Fairness and Bias Mitigation

Shifting from resume-based to skill-based screening reduces several well-documented sources of hiring bias, including name-based discrimination, educational prestige bias, and formatting-related inconsistencies [3]. However, skill assessments can introduce new biases if question content favors

candidates from particular linguistic, geographic, or educational backgrounds.

The proposed system mitigates these risks through question content review workflows, difficulty calibration across demographic segments, and score distribution monitoring. Where demographic data is available, score distributions are analyzed per group to identify potential disparate impact. Ongoing fairness auditing is identified as a critical operational requirement, particularly as the system is deployed across diverse candidate pools and role types.

VIII. CONCLUSION AND FUTURE WORK

This paper presented an AI-assisted skill-based candidate evaluation system that replaces resume-centric screening with a unified, performance-driven assessment pipeline. The system integrates candidate profiling, competency-mapped skill matching, dynamic assessment generation, multi-modal automated evaluation, and weighted ranking into a single end-to-end framework.

Experimental evaluation on simulated data demonstrated substantial improvements over a conventional ATS baseline: Spearman rank correlation improved from 0.61 to 0.91, the false-positive shortlist rate fell from 34% to 12%, and top-quintile precision increased from 41% to 78%. The modular four-layer architecture ensures scalability and enables incremental enhancement of individual components without pipeline disruption.

Several directions for future work are identified. First, the NLP evaluation component will incorporate SBERT [6] to replace TF-IDF with semantic sentence embeddings robust to paraphrasing and domain-specific vocabulary variation, expected to improve descriptive response scoring accuracy. Second, supervised learning will be applied to optimize scoring weights from post-hire performance data, replacing the current heuristic weight configuration with a data-driven approach. Third, automated competency model generation from job description corpora using ontology learning [7] will reduce manual authoring overhead and enable broader role coverage. Fourth, the assessment integrity mechanisms will be extended with passive proctoring signals to further verify response authenticity at scale. Finally, a longitudinal study tracking post-hire outcomes for candidates evaluated by the proposed system will be conducted to validate predictive validity and provide ground-truth signal for continuous model improvement.

As AI-assisted hiring becomes more prevalent across industries, frameworks that prioritize demonstrated skill over document-based proxies represent a critical direction for building both efficient and equitable recruitment pipelines. The proposed system provides a concrete, implementable foundation for this direction.

Acknowledgment
The authors acknowledge the beta users of the proposed platform whose feedback informed the design and evaluation of the system.

REFERENCES

1. Jobscan, “2023 Applicant Tracking System Usage Report,” Jobscan Inc., 2023. [Online]. Available: https://www.jobscan.co/blog/fortune-500-ats/
2. A. Rios et al., “Resume2Vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching,” Electronics, vol. 14, no. 4, p. 794, Feb. 2025. doi: 10.3390/electronics14040794
3. D. F. Mujtaba and N. R. Mahapatra, “Ethical considerations in AI-based recruitment,” in Proc. 2019 IEEE Int. Symp. Technology and Society (ISTAS), Nov. 2019, pp. 1–7. doi: 10.1109/ISTAS48451.2019.8937920
4. I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, “Automated analysis and prediction of job interview performance,” IEEE Trans. Affective Computing, vol. 9, no. 2, pp. 191–204, 2018. doi: 10.1109/TAFFC.2016.2614299
5. G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988. doi: 10.1016/0306-4573(88)90021-0
6. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. EMNLP-IJCNLP 2019, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410
7. T. Petrican et al., “Ontology-based skill matching algorithms,” in Proc. 2017 IEEE Int. Conf. Intelligent Computer Communication and Processing (ICCP), Sep. 2017, pp. 205–211. doi: 10.1109/ICCP.2017.8117005
8. G. Kurdi et al., “A systematic review of automatic question generation for educational purposes,” Int. J. Artif. Intell. Educ., vol. 30, no. 1, pp. 121–204, 2020. doi: 10.1007/s40593-019-00186-y
9. B. Das et al., “Automatic question generation and answer assessment: a survey,” Res. Pract. Technol. Enhanc. Learn., vol. 16, no. 1, p. 5, 2021. doi: 10.1186/s41039-021-00151-1