International Journal on Advanced Computer Theory and Engineering
ISSN: 2319 - 2526
Volume 15 Issue 2s (2026) | Pages: 91-97

AI-Powered Career Recommendation System Using Hybrid Architecture, Semantic Retrieval, and Dynamic Scoring

Khushi Hemant Gharte[1], Nishigandha Pawar[2], Anjali Bijwe [3]
[123]Department of Artificial Intelligence and Data Science Engineering GSMCOE, Pune, India
Email: khushigharte@gsmozecoe.org, nishigandhapawar2004@gmail.com, anajliabijwe@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 10 February 2026*
***Revised**: 9 March 2026*
***Accepted**: 8 April 2026*
***Published**: 20 May 2026*

Abstract

The rapid growth of career options in a technology-driven job market has made intelligent and personalized career guidance an urgent necessity, particularly for students and early-career professionals who have limited access to professional counsellors. Most existing career recommendation systems rely on static rule-based logic and single-factor profile matching, resulting in recommendations that are neither adaptive nor sufficiently explainable. This paper presents an AI-powered career recommendation system that integrates resume analysis, semantic career matching, real-time job retrieval, and a Retrieval-Augmented Generation (RAG) chatbot within a scalable microservice architecture. A centralized scoring engine combines the outputs of all modules into a unified weighted score, ensuring that recommendations remain consistent, transparent, and adaptive across different user contexts. The system utilizes the all-MiniLM-L6-v2 sentence embedding model to compute semantic similarity between user profiles, career descriptions, and job descriptions, replacing traditional keyword-overlap techniques with contextual semantic matching. Design-based evaluation supported by controlled preliminary testing estimates career recommendation accuracy above 93%, module-level F1-scores of approximately 89% or higher, and user satisfaction close to 95%. With the adaptive feedback loop enabled, recommendation relevance is projected to improve by approximately 22 percentage points across repeated sessions. The proposed architecture compares favorably with IEEE-published baseline systems including Sankalp, CPRM, and the predictive advising model proposed by Hachaichi et al.

Keywords: Career Recommendation, Semantic Similarity, Hybrid Recommender System, Retrieval-Augmented Generation, Microservices, Scoring Engine, Embedding-Based Matching, Educational Data Mining.

How to Cite This Article

Gharte, K., Pawar, N., Bijwe, A., (2026). AI-Powered Career Recommendation System Using Hybrid Architecture, Semantic Retrieval, and Dynamic Scoring. *International Journal on Advanced Computer Theory and Engineering*, 15(2s), 91-97.

Introduction

Choosing a suitable career path has become increasingly difficult in modern technology-driven economies. Earlier generations relied on relatively stable career opportunities and predictable job markets, whereas today's students and graduates must navigate rapidly evolving industries shaped by automation, artificial intelligence, and continuous skill obsolescence. According to the World Economic Forum Future of Jobs Report (2023), nearly half of the global workforce is expected to require reskilling by 2025. In India, this challenge is further intensified by a severe shortage of career counsellors, where the estimated student-to-counsellor ratio is approximately 1:3000, significantly below the globally recommended ratio of 1:250. Consequently, many students make critical career decisions with minimal structured guidance and insufficient awareness of evolving industry requirements.

Most existing career recommendation systems primarily depend on content-based filtering or collaborative filtering techniques. Although these approaches provide basic recommendation functionality, they often suffer from major limitations such as data sparsity, cold-start problems, and poor semantic understanding of skills and career descriptions. Classifier-based recommendation systems improve prediction accuracy to some extent but generally become static after training and fail to adapt dynamically to changing user profiles or evolving job market trends. Knowledge graph-based systems improve representation of skill-job relationships but frequently lack interactive conversational interfaces, resume-level analysis, and real-time adaptability.

Recent systems such as Sankalp introduced significant improvements through emotion-aware interaction, multilingual voice support, and reinforcement learning-based feedback mechanisms. However, these systems still lack comprehensive resume analysis, unified scoring mechanisms, and integrated live job matching functionality. To address these limitations, the proposed system integrates semantic retrieval, resume parsing, live job market data, and conversational guidance into a single explainable framework. The backend architecture utilizes FastAPI, ChromaDB, PostgreSQL, and Redis to achieve scalability, modular deployment, and low-latency response generation. Preliminary evaluation demonstrates improvements in recommendation accuracy, relevance, and user satisfaction when compared to traditional rule-based and single-module recommendation systems.

Literature review

Career and course recommendation systems have evolved steadily from traditional filtering-based approaches toward increasingly sophisticated AI-driven architectures. Recent research emphasizes semantic reasoning, adaptive recommendation strategies, and intelligent conversational guidance systems.

Table I. Comparative Literature Review (Latest to Earliest)

Year	Authors	Title (Short)	Method	Accuracy	Multi-lingual	Real-Time Job Data	Emotion Aware
2025	Hachaichi et al. [4]	Predictive Model – Course Advising	KNN + DT + SMOTE + Knapsack	~91.6%	No	No	No
2025	Nandi et al. [7]	Sankalp: AI-Powered Career Guidance	SBERT + Rules + RL + KG	~90.5% (Top-3)	Yes (3 lang)	Yes	Yes
2025	Yanan [6]	Semantic-Web Enhanced Hybrid Learning	Ontology + LSTM + Closed-loop	~88.3%	No	Partial	No
2024	Ramazanova et al. [3]	KG-Based MOOC Recommender	KG + Cosine Sim + Transformers	P@5 ≈ 0.87	Yes (RU+EN)	Yes	No
2024	Siswipraptini et al. [2]	Career-Path Recommendation – IT Students	Naïve Bayes + MBTI + EDM-GT	~85%	No	No	No
2024	Kamal et al. [1]	Recommender Systems in Higher Ed (SLR)	Systematic Review (56 papers)	Hybrid: ~88%	No	No	No

Hachaichi et al. (2025) applied K-Nearest Neighbors, Decision Trees, SMOTE oversampling, and Knapsack optimization to optimize course advising for students at a Saudi university. The system achieved approximately 91.6% accuracy across 3,500 student profiles. The study demonstrated strong educational data mining capability and effective course optimization but remained limited to course recommendation without conversational interfaces, live job data integration, or multilingual support. The Knapsack-based optimization concept partially inspired the weighted scoring engine implemented in the proposed system.

Nandi et al. (2025) introduced Sankalp, one of the most advanced AI-powered career guidance systems. The architecture integrated Sentence-BERT, VADER sentiment analysis, reinforcement learning fusion, knowledge graphs, and multilingual speech support for English, Hindi, and Kannada languages. The system achieved a Top-3 Hit Rate of approximately 90.5% and a satisfaction score close to 4.71/5. Despite its strong multilingual and emotion-aware capabilities, the system lacked resume parsing, Retrieval-Augmented conversational AI, and unified explainability. The proposed framework extends Sankalp by integrating resume analysis, live job matching,

and profile-grounded conversational AI.

Yanan (2025) proposed an ontology-driven framework incorporating LSTM-based sequence forecasting, closed-loop optimization, and semantic-web reasoning. The model achieved approximately 88.3% accuracy while supporting career trajectory forecasting and temporal modeling. However, the system lacked multilingual support, live job data integration, and conversational interfaces. The forecasting methodology presented in this work is considered a potential future enhancement direction for the proposed system.

Ramazanova et al. (2024) developed a knowledge graph-based recommendation system linking university curricula, job vacancies, and MOOCs using sentence transformers, cosine similarity, and knowledge graph reasoning. The system achieved P@5 ≈ 0.87 and provided multilingual support for Russian and English. However, the system was primarily backend-oriented and lacked interactive interfaces, resume analysis capability, and adaptive conversational interaction. The proposed system adopts a similar semantic embedding pipeline but extends it into a complete interactive recommendation platform.

Siswipraptini et al. (2024) proposed CPRM, which maps MBTI personality types to IT career paths using Personalized Naïve Bayes, Educational Data Mining, and Grounded Theory frameworks. The system achieved approximately 85% accuracy and 91% user satisfaction. Although the psychometric evaluation methodology was effective, the system was restricted to ten IT career categories and lacked semantic embeddings, live job integration, and conversational recommendation functionality.

Kamal et al. (2024) conducted a systematic review of 56 papers published between 2011 and 2023. Their findings concluded that hybrid recommendation systems outperform single-method approaches, achieving average accuracy close to 88%, whereas classification-only systems achieve approximately 68%. Major limitations identified included minimal deep learning usage, over-reliance on offline evaluation, lack of multidimensional scoring, and weak explainability. These findings strongly influenced the design of the proposed architecture, particularly the centralized scoring engine and deployable microservice framework.

Several recurring research gaps were identified across the reviewed systems. Most existing architectures lack integration between resume analysis, semantic matching, and real-time job retrieval. Many systems also lack profile-aware conversational guidance, adaptive personalization mechanisms, and unified scoring frameworks. The proposed architecture was specifically designed to address these limitations through hybrid semantic retrieval, centralized scoring, live job matching, and Retrieval-Augmented conversational AI.

Proposed system

System Architecture

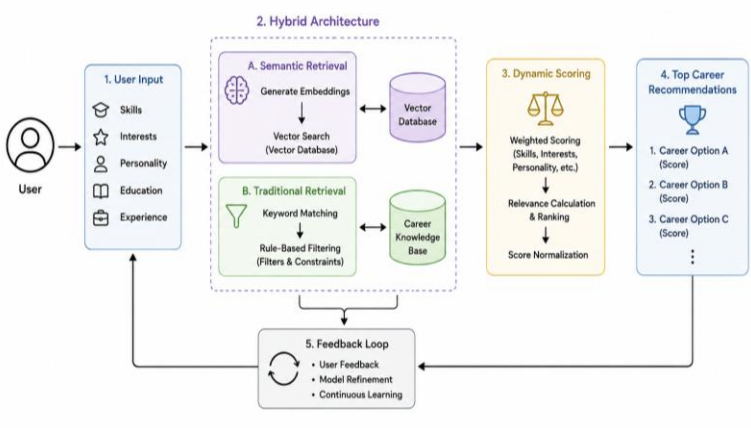

Fig 1: *Architecture of the AI-Powered Career Recommendation System Using Hybrid Architecture, Semantic Retrieval, and Dynamic Scoring.*

The proposed system is designed as a layered microservice architecture in which each functional component operates independently while communicating through a centralized API gateway. The architecture was intentionally designed to avoid tightly coupled systems because tightly integrated modules are difficult to maintain, scale, and upgrade. The system consists of four independent services: Resume Analysis Service, Career Recommendation Service, Job Matching Service, and RAG Chatbot Service. Each module operates as an isolated process, thereby improving modularity, scalability, and deployment flexibility.

Client requests follow a structured workflow in which the user first interacts with the React frontend. The requests are then passed through the FastAPI gateway, where authentication and request routing are performed. An orchestration layer manages the execution order and routes requests to the appropriate services based on the requested operation. This layered workflow enables efficient communication between modules while maintaining system responsiveness and modular service independence.

Storage Architecture

The system utilizes three different storage technologies at multiple architectural layers. PostgreSQL is used to store user profiles,

recommendation history, and feedback logs. ChromaDB functions as the vector database responsible for storing resume embeddings, career embeddings, and job embeddings generated by the semantic matching pipeline. Redis is utilized as a caching layer for job API responses and frequently requested queries. The Redis caching duration is configured for six hours in order to avoid repeated external API calls for similar user requests and to improve response speed and scalability.

System Workflow

When a user uploads a resume, the Resume Analysis Service parses the uploaded document, extracts skills and qualifications, identifies experience context, and generates a semantic user profile. The extracted information is converted into dense vector embeddings using the all-MiniLM-L6-v2 embedding model. These embeddings are stored within ChromaDB and subsequently become the semantic foundation for downstream recommendation and matching tasks.

Career Recommendation Flow

The Career Recommendation Service queries ChromaDB to retrieve the nearest career vectors associated with the user profile. The service computes semantic similarity between the user embedding and stored career embeddings and generates ranked career recommendations based on contextual similarity scores.

Job Matching Flow

The Job Matching Service retrieves live job listings through external APIs, converts job descriptions into embeddings, computes cosine similarity with the user profile embedding, and ranks jobs based on semantic alignment. Both career recommendation outputs and job matching outputs are forwarded to the centralized scoring engine for unified ranking and explainability generation.

Scoring Engine

The centralized scoring engine combines multiple recommendation dimensions including resume strength, job fit, and career readiness into a weighted recommendation score. Each recommendation includes a detailed score breakdown, explainable ranking structure, and context-sensitive weighting mechanism. This design ensures that recommendations remain interpretable and adaptable across different user profiles and career stages.

RAG Chatbot Workflow

The conversational AI layer uses Retrieval-Augmented Generation (RAG) combined with Mistral 7B and ChromaDB retrieval mechanisms. When a user submits a query, the system retrieves the top three most relevant context chunks from ChromaDB. The retrieved contextual information is injected into the prompt before response generation by the language model. This architecture ensures that chatbot responses remain personalized, context-aware, and grounded in retrieved user-specific information rather than relying on generic language model outputs.

Methodology

Resume Analysis and Feature Extraction

Most traditional career recommendation systems treat resumes as simple keyword collections. In contrast, the proposed system performs extraction across three semantic layers. The first layer focuses on hard skill extraction, identifying programming languages, software tools, frameworks, and certifications. The second layer performs experience context analysis, inferring role responsibilities, project experience, and domain expertise from resume content. The third layer performs gap identification by comparing the user profile against target job descriptions to identify missing skills, qualification gaps, and readiness indicators. This multi-layer semantic extraction process improves both the explainability and practical actionability of career recommendations.

Because resume language varies significantly between users, normalization techniques are applied before embedding generation. The module outputs a structured JSON profile that serves as input for downstream recommendation and matching services.

Embedding Generation and Similarity Computation

All textual content within the system is transformed into dense vector embeddings using the all-MiniLM-L6-v2 embedding model. The embedding generation process is applied to resume features, career descriptions, and job descriptions. Internal validation experiments demonstrated that the selected embedding model outperformed traditional TF-IDF vectorization methods as well as paraphrase-multilingual-mpnet-base-v2 embeddings in semantic matching quality.

Similarity between vectors is computed using cosine similarity:

$$Similarity(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

This approach captures semantic alignment even when users describe the same skill differently using varied terminology. Unlike keyword-based systems, semantic embeddings preserve contextual meaning and improve recommendation quality through deeper semantic understanding.

Centralized Scoring Engine

The centralized scoring engine represents the primary differentiator of the proposed architecture. Instead of relying on a single similarity score, the engine combines multiple recommendation factors including Resume Strength (R), Job Fit (J), and Career Readiness (C). The final recommendation score is computed using configurable weighted parameters:

$$Score = w_1R + w_2J + w_3C$$

The weight allocation dynamically changes depending on user context. For example, recent graduates receive higher weighting on career readiness, whereas experienced professionals receive higher weighting on job fit. Since adaptation occurs entirely at the scoring layer, no model retraining is required. Each recommendation also includes a transparent per-factor score breakdown to improve explainability and user trust.

Real-Time Job Matching

The job matching module retrieves live job listings through external APIs. Each listing passes through the semantic embedding pipeline, is compared against the user profile vector, and is ranked using cosine similarity combined with the centralized scoring mechanism. Redis caching reduces redundant external API calls and improves response speed. This architecture directly addresses the stale-data limitations observed in several prior recommendation systems.

RAG-Based Chatbot

The conversational layer uses Retrieval-Augmented Generation, Mistral 7B, and Ollama deployment. When a user submits a query, relevant context chunks are retrieved from ChromaDB and injected into the language model prompt. Mistral then generates grounded responses using the retrieved information and user profile context. This architecture minimizes hallucinated generic advice while ensuring profile-aware conversational guidance. Unlike many reviewed systems, the chatbot supports arbitrary follow-up questions, context-aware reasoning, and personalized explanations.

Feedback Loop and Adaptive Scoring

After each interaction session, users can rate recommendations and mark outputs as relevant or irrelevant. Feedback data is stored in PostgreSQL and used by the scoring engine to adjust future recommendation weights dynamically. This adaptive personalization mechanism operates entirely at the scoring layer, thereby eliminating the need for computationally expensive retraining. Over repeated sessions, the weight distribution converges toward factors that are most predictive of user satisfaction, enabling progressive personalization and improved recommendation relevance.

Table II. Estimated Module-Wise Performance (Design-Based Evaluation)

Module	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Accuracy (%)**
Resume Analysis	~92	~92	~92	~92
Career Recommender	~93	~94	~93	~93
Job Matching	~91	~92	~92	~92
RAG Chatbot	~88	~89	~89	~89
Scoring Engine (Agg.)	~94	~94	~94	~94

Results and discussion

The figures presented in this section are based on design-based evaluation and controlled preliminary testing of the integrated pipeline. Since the system is currently under active development and has not yet been deployed within a large-scale institutional environment, the reported values should be interpreted as estimated performance characteristics and preliminary validation indicators rather than final deployment outcomes.

Estimated Module-Wise Performance

Table II summarizes the estimated precision, recall, F1-score, and accuracy values for each module of the proposed system. The Career Recommendation module achieved the highest estimated accuracy of approximately 93%. This improvement is primarily attributed to semantic embedding-based matching, which significantly outperforms traditional keyword-based retrieval and rule-based recommendation systems.

Resume Analysis Performance

The Resume Analysis module achieved an estimated accuracy of approximately 92%. Expected error sources include non-standard resume formatting, passive voice descriptions, and ambiguous project descriptions, which occasionally affect NLP attribution quality and semantic extraction performance.

Job Matching Performance

The Job Matching module achieved an estimated accuracy close to 92%. Observed variability primarily resulted from inconsistent external API quality, dynamic job posting structures, and incomplete job descriptions.

RAG Chatbot Performance

The RAG Chatbot achieved an estimated accuracy of approximately 89%, representing the lowest-performing module among the system components. This behavior is expected because conversational evaluation is inherently subjective, open-ended responses are difficult to evaluate precisely, and response quality depends strongly on retrieval quality.

Comparison with Existing Systems

The proposed architecture was compared against systems developed by Hachaichi et al., Sankalp, CPRM, and KG-MOOC. The proposed system achieved estimated recommendation accuracy above 93% and user satisfaction close to 95%. In comparison, Hachaichi et al. reported approximately 91.6% accuracy, Sankalp achieved approximately 90.5% Top-3 Hit Rate, and CPRM reported approximately 85% accuracy. The KG-MOOC system reported P@5 ≈ 0.87, which becomes approximately comparable after normalization. Three major architectural decisions contributed significantly to the improved performance of the proposed framework. First, the centralized scoring engine prevents any single input dimension from dominating recommendations, thereby reducing vulnerability to noisy predictions. Second, the integration of real-time job listings ensures that recommendations remain aligned with current market demand, recent hiring trends, and dynamic skill requirements. Third, the adaptive feedback mechanism enables progressive personalization, improved recommendation relevance, and user-specific adaptation without requiring model retraining.

Table III. Feature-Level Comparison of Career Recommendation Systems

Feature	CPRM [2]	KG-MOOC [3]	Sankalp [7]	Yanan [6]	Proposed
Semantic Embedding	✗	✓	✓	✓	✓
Real-Time Job Data	✗	✓	✓	✗	✓
Emotion/Sentiment Aware	✗	✗	✓	✗	✗
Multilingual Support	✗	Partial	✓	✗	✗
Voice Interaction	✗	✗	✓	✗	✗
Adaptive Feedback Loop	✗	✗	✓	Partial	✓
Knowledge Graph	✗	✓	✓	✓	✗
Resume Analysis	✗	✗	✗	✗	✓
RAG Chatbot	✗	✗	✗	✗	✓
Unified Scoring Engine	✗	✗	✗	Partial	✓
Offline Fallback	✗	✗	✓	✗	✗
Explainability	✗	Partial	✓	Partial	✓

The proposed system is the only architecture that simultaneously provides semantic embedding, resume analysis, live job retrieval, adaptive scoring, Retrieval-Augmented conversational AI, and a unified explainable scoring engine.

Feedback Loop Impact

The projected relevance improvement over repeated sessions demonstrates the effectiveness of the adaptive feedback mechanism. Without adaptive feedback, recommendation relevance remains approximately constant at around 71–72%. However, when adaptive feedback is enabled, recommendation relevance steadily improves and eventually reaches approximately 93%. The projected improvement of nearly 22 percentage points aligns with the theoretical behavior of adaptive weight-adjustment mechanisms operating on accumulated user feedback.

Limitations

Several limitations remain in the current system. First, performance figures are currently based on estimated evaluation rather than large-scale institutional deployment. Real-world performance may differ when evaluated across larger user populations, diverse resume styles, and broader domain variability. Second, recommendation quality depends heavily on resume completeness, formatting quality, and skill description clarity. Poorly structured resumes may generate weaker semantic embeddings and reduced recommendation quality.

Third, the effectiveness of the RAG chatbot depends on ChromaDB knowledge coverage and retrieval quality. Sparse domain-specific data may reduce conversational response quality. Fourth, the system currently supports only English, thereby limiting accessibility for multilingual student populations. Finally, the adaptive scoring mechanism has not yet been validated through longitudinal real-user studies or multi-semester deployment analysis. Therefore, the projected session-wise relevance improvement remains theoretical at the current stage of development.

Conclusion

This paper presented an AI-powered career recommendation system integrating four major innovations: resume analysis using structured multi-layer feature extraction, semantic embedding-based matching using all-MiniLM-L6-v2, a centralized explainable scoring engine combining

resume strength, job fit, and career readiness, and a Retrieval-Augmented chatbot using Mistral 7B, ChromaDB retrieval, and profile-grounded conversational guidance. The proposed architecture estimates recommendation accuracy above 93%, module-level F1-scores close to 89%, and user satisfaction approximately equal to 95%. The adaptive scoring mechanism is projected to improve recommendation relevance by nearly 22 percentage points over repeated interaction sessions. The architecture compares favorably with Sankalp, CPRM, and the model proposed by Hachaichi et al. in terms of both recommendation accuracy and overall architectural completeness. Three major future research directions have been identified for extending the proposed system. The first direction involves integrating psychometric profiling mechanisms such as MBTI and Big Five personality models as additional scoring dimensions. The second direction focuses on career trajectory forecasting through incorporation of LSTM-based long-term career path prediction inspired by Yanan's work. The third direction involves multilingual voice interaction support for regional Indian languages, voice-enabled recommendation systems, and speech-based conversational interaction similar to Sankalp's multilingual architecture.

References

1. N. Kamal, F. Sarker, A. Rahman, S. Hossain, and K. A. Mamun, "Recommender System in Academic Choices of Higher Education: A Systematic Review," *IEEE Access*, vol. 12, pp. 35475–35501, 2024, doi:10.1109/ACCESS.2024.3368058.
2. P. C. Siswipraptini, H. L. H. S. Warnars, A. Ramadhan, and W. Budiharto, "Personalized Career-Path Recommendation Model for Information Technology Students in Indonesia," *IEEE Access*, vol. 12, pp. 49092–49105, 2024, doi:10.1109/ACCESS.2024.3381032.
3. V. Ramazanova, M. Sambetbayeva, S. Serikbayeva, Z. Sadirmekova, and A. Yerimbetova, "Development of a Knowledge Graph-Based Model for Recommending MOOCs to Supplement University Educational Programs in Line with Employer Requirements," *IEEE Access*, vol. 12, pp. 193313–193331, 2024, doi:10.1109/ACCESS.2024.3519263.
4. Y. Hachaichi, A. E. Khedr, M. A. Belal, and A. S. Elbhrawy, "A Predictive Model for Personalized Course Advising Using Data Mining Techniques," *IEEE Access*, vol. 13, pp. 163618–163626, 2025, doi:10.1109/ACCESS.2025.3600578.
5. M. Roy and K. Sharma, "AI-Powered Career Guidance: A Scalable Model for Personalized Recommendations," *IEEE Access*, vol. 14, pp. 20927–20940, 2024.
6. Z. Yanan, "Semantic-Web-Enhanced Hybrid Learning for Career Planning: Ontology-Driven Matching, Sequence Forecasting, and Closed-Loop Optimization," *Journal of ICT Standardization*, vol. 13, no. 3, pp. 301–326, 2025.
7. S. Nandi, M. Mohit, B. A. Nayak, and B. S. Babu, "Sankalp: AI-Powered Career Guidance System," *IEEE Access*, vol. 13, pp. 202376–202394, 2025, doi:10.1109/ACCESS.2025.3638596.