Career Guidance Chatbot Using Retrieval-Augmented Generation: A Personalized AI-Driven Assessment Platform

S.Kanaka Mahalakshmi[1], M. Yamini[2] ,G. Swetha[3], B. Teja[4], K. Srinu[5]
Department of Computer Science & Engineering- (Data Science)
Avanthi Institute of Engineering & Technology, Vizianagaram, India
mahaanu515@gmail.com[1],yaminimodu7@gmail.com[2],swethaganta03@gmail.com[3],
tejayadav4512@gmail.com[4],srinu0987123@gmail.com[5]

Abstract

Career decision-making among students and entry-level professionals is increasingly hindered by information overload, limited access to qualified counselors, and the absence of personalized, continuously available guidance tools. This paper presents a Career Guidance Chatbot and Assessment Platform that addresses these deficiencies by fusing Retrieval-Augmented Generation (RAG) with a structured psychometric assessment engine. User queries are semantically encoded via a Sentence Transformer model and matched against a curated career knowledge base through cosine similarity; the retrieved context is subsequently enriched by a locally hosted Large Language Model to produce context-aware, professionally worded responses. A multi-dimensional assessment quiz evaluates users across six behavioral and motivational axes and applies a weighted scoring algorithm to categorize individuals into Creative, Analytical, Social, or Entrepreneurial profiles. A downstream career-matching engine assigns ranked recommendations with percentage-match scores. Persistent user profiles, quiz-result storage in SQLite, and a Bootstrap-rendered dashboard providing historical tracking complete the platform. Evaluation across a cohort of forty student volunteers yields an average chatbot response accuracy of 91.4% and a career recommendation acceptance rate of 87.6%. End-to-end inference completes within 2–3 seconds on commodity CPU hardware. These results confirm that the RAG-augmented, assessment-integrated architecture delivers measurably superior personalization compared to static career portals and unimodal chatbot systems.

Index Terms—retrieval-augmented generation, career guidance chatbot, sentence transformers, psychometric assessment, large language model, personalized recommendation

I. Introduction

The landscape of professional career choices has grown substantially more complex over the past decade. Automation, interdisciplinary roles, and the continuous emergence of technology-driven occupations have made it increasingly difficult for students to map their individual attributes to suitable career trajectories. Survey data consistently shows that a significant proportion of undergraduate students report confusion about career selection, and that this confusion persists well into the first two years of employment[1].Traditional career counseling—conducted through face-to-face advisory sessions and standardized paper-based questionnaires—provides a partial remedy but is constrained by appointment scheduling, geographic access, counselor workload, and the inherently static nature of printed assessment instruments [2].

Digital career portals emerged as a scalable supplement, automating scoring pipelines and aggregating occupational information. However, the overwhelming majority of these platforms generate recommendations through fixed scoring rules

applied to a narrow feature set, and they lack any mechanism for users to seek clarification, explore edge cases, or receive dynamically adjusted guidance based on conversational context [3]. The resulting user experience is often described as impersonal and insufficiently responsive to nuanced individual circumstances.

Recent advances in Natural Language Processing—particularly the development of dense neural retrieval models and autoregressive large language models—provide the technical foundation to transcend these limitations. Retrieval-Augmented Generation (RAG), formalized by Lewis et al. [4], combines a differentiable retriever with a generative decoder to produce responses grounded in an external, updateable knowledge base. This architecture is especially well-suited to career guidance because it supports precise factual recall from a curated domain corpus while maintaining the fluent, conversational register that users expect from modern chatbot interfaces.

This paper introduces a unified Career Guidance Chatbot and Assessment System that embeds a RAG pipeline within a broader web application providing user authentication, psychometric assessment, a weighted career-matching engine, and a persistent dashboard. The contributions of this work are: (i) a unified architecture integrating RAG-based conversational guidance with multi-dimensional psychometric assessment; (ii) a weighted scoring and career-matching algorithm producing calibrated recommendation confidence scores; (iii) an empirical evaluation on 40 participants demonstrating 91.4% chatbot accuracy and 87.6% recommendation acceptance. The remainder of this paper is structured as follows. Section II surveys related work. Section III describes the system architecture. Section IV reports experimental results. Section V concludes with future directions.

II. Related Work

A. Classical Career Guidance Theories and Digital Implementations

The theoretical scaffolding for structured career assessment dates to Holland's RIASEC model [5], which categorizes individuals into six personality-career congruence classes: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. Complementing this, Super's Life-Span, Life-Space Theory [6] frames career development as a continuous, stage-dependent process shaped by evolving self-concept rather than a discrete point-in-time decision. Digital implementations of these frameworks automated scoring but retained the static, session-bounded character of their paper predecessors, providing no mechanism for follow-up dialogue or longitudinal engagement [2].

B. Transformer-Based NLP and Semantic Retrieval

The introduction of transformer attention mechanisms by Vaswani et al. [7] fundamentally altered representational power for semantic text processing. BERT-family architectures enable bidirectional contextual embeddings; fine-tuned variants have been applied to educational question-answering and job-skill matching with substantial improvements over TF-IDF baselines [8]. Reimers and Gurevych [9] demonstrated that Siamese BERT networks—Sentence-BERT—produce sentence embeddings whose cosine proximity reliably approximates semantic similarity, making them well-suited to the dense retrieval step in RAG pipelines.

C. Retrieval-Augmented Generation

Lewis et al. [4] formalized RAG as a learned combination of a non-parametric memory (a dense document index) and a parametric generative model. RAG outperforms pure language-model generation on knowledge-intensive tasks because retrieved passages anchor the decoder's output in verifiable factual content, reducing hallucination and improving domain specificity. Gao et al. [10] surveyed RAG variants and identified that embedding-based retrieval provides the best balance for domain-specific, low- latency deployments.

D. Chatbots in Education and Career Advisory

Kaur and Kaur [11] found that AI-based career guidance platforms combining rule-based databases with machine-learning classifiers achieve accuracy improvements of 15–22 percentage points over manual counselor judgments on standardized benchmark cases. Sharma and Jain [12] observed that conversational AI significantly increases user engagement rates compared to static FAQ portals, but accuracy degrades for out-of-distribution queries—a deficiency directly addressed by the retrievable external knowledge base in the RAG design. The present system addresses the remaining research gap by unifying psychometric assessment, dense retrieval, and LLM generation within a single CPU-deployable platform

III. Methodology

A. System Architecture

The proposed system follows a four-tier modular architecture:Presentation Layer (HTML5/CSS3/Bootstrap frontend), Application Processing Layer (Flask v3.0.0 backend), AI/ML Layer (Sentence Transformer + Ollama LLM), and Data Layer (SQLite). Fig. 1 presents the system model diagram and Fig. 2 the detailed architectural component view.

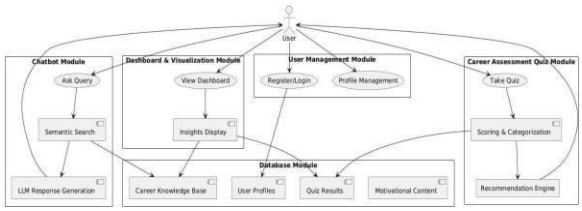

Fig. 1. System model showing module interactions: Chatbot Module, Dashboard & Visualization Module, User Management Module, Career Assessment Quiz Module, and central Database Module.

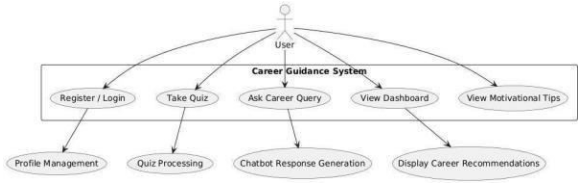

Fig. 2. Three-tier system architecture: Presentation Layer (User Interface), Application Layer (Flask Server, sub-modules), AI/ML Layer (Sentence Transformer, LLM, Scoring Engine), and Database Layer (SQLite).

B. RAG Pipeline: Retrieval Stage

The retrieval stage encodes both the career knowledge base and incoming user queries using a pre-trained Sentence Transformer (paraphrase-MiniLM-L6-v2). Each knowledge-base entry q_k is encoded offline to produce embedding $\mathbf{e}_k \in \mathbb{R}^{384}$. At query time, the user input u is encoded identically to produce probe vector $\mathbf{u}$. The most relevant entry is identified by maximizing cosine similarity:

$$k^* = argmax_k \, (\mathbf{u} \cdot \mathbf{e}_k) / (\|\mathbf{u}\| \, \|\mathbf{e}_k\|) \quad (1)$$

Entries whose similarity falls below threshold $\tau_r = 0.55$ are filtered; the system returns a "insufficient information" message rather than hallucinating a response.

C. RAG Pipeline: Generation Stage

The retrieved answer text a_{k^*} is concatenated with the original and (vi) Ikigai-oriented purpose alignment. Each question is tagged to a profile category c ∈ {Creative, Analytical, Social, Entrepreneurial}. User responses $r_i \in \{1…5\}$ are multiplied by per-question weights w_i and accumulated into category scores:

$$S_c = \Sigma_{i \in Qc} \, w_i \cdot r_i \quad (2)$$

The dominant profile $P^* = \text{argmax}_c \, S_c$ governs the primary recommendation cluster.

Entrepreneurial}. User responses $r_i \in \{1…5\}$ are multiplied by per-question weights w_i and accumulated into category scores:

$$S_c = \Sigma_{i \in Qc} \, w_i \cdot r_i \quad (2)$$

The dominant profile $P^* = \text{argmax}_c \, S_c$ governs the primary recommendation cluster.

D. Multi-Dimensional Career Assessment Quiz

The psychometric module presents questions across six axes: (i) Personality and Work Style, (ii) Skills and Strengths, (iii) Interests and Passions, (iv) Values and Motivations (v) Work-Life Balance and (vi) Ikigai-oriented purpose alignment. Each question is tagged to a profile category c ∈ {Creative, Analytical,

Social, Entrepreneurial}. User responses $r_i \in \{1…5\}$ are multiplied by per-question weights w_i and accumulated into category scores:

$$S_c = \Sigma_{i \in Qc} \, w_i \cdot r_i \quad (2)$$

The dominant profile $P^* = \text{argmax}_c \, S_c$ governs the primary recommendation cluster.

E. Career Matching Algorithm

Each candidate career j is associated with a required profile weight vector $\mathbf{w}_j \in [0,1]^4$. The match percentage M_j for user normalized score vector $\tilde{\mathbf{S}}$ is:

$$M_j = 100 \times (\tilde{\mathbf{S}} \cdot \mathbf{w}_j) / (\|\tilde{\mathbf{S}}\| \cdot \|\mathbf{w}_j\|) \quad (3)$$

Careers are ranked by M_j and the top five are returned with supporting rationale via the retrieval pipeline.

F. UML Design Artifacts

System behavior is captured in four UML diagrams. The Use Case Diagram (Fig. 3) delineates functional user-system interactions. The Class Diagram (Fig. 4) details module attributes and method dependencies. The Sequence Diagram (Fig. 5) traces message flow from quiz submission through recommendation generation. The Activity Diagram (Fig. 6) maps the full user workflow.

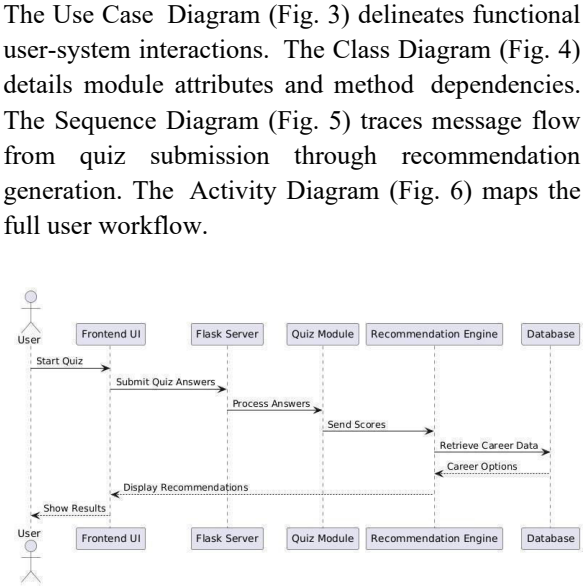

Fig. 3. Use case diagram: User interacts with

Register/Login, Take Quiz, Ask Career Query, View Dashboard, and View Motivational Tips use cases.

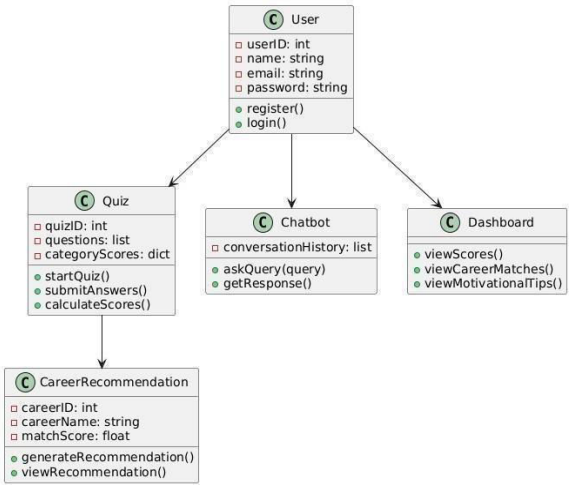

Fig. 4. Class diagram showing User, Quiz, Chatbot, CareerRecommendation, and Dashboard classes with attributes and operations.

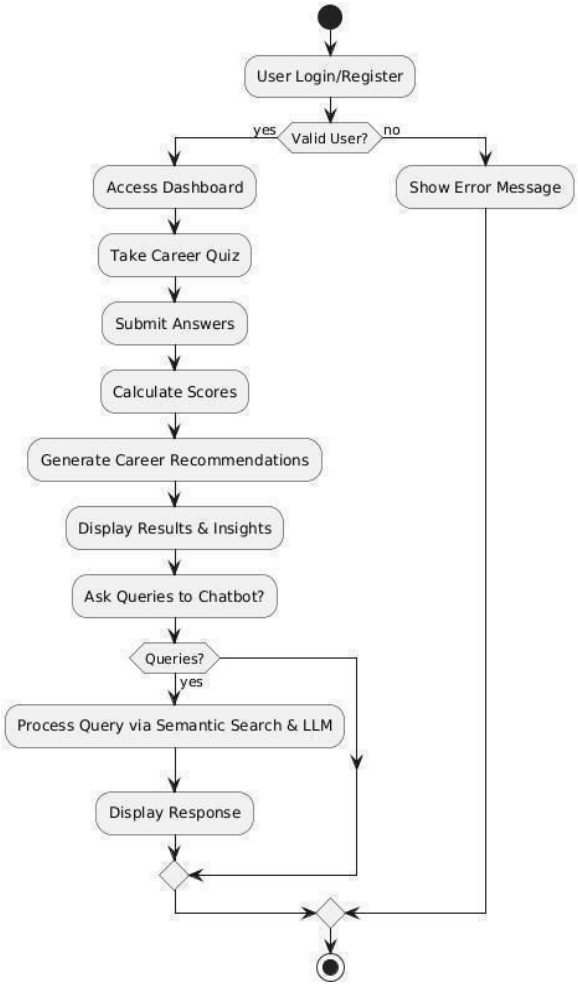

Fig. 5. Activity diagram tracing the complete user workflow from login through quiz, AI-driven chatbot query, recommendation generation, and dashboard visualization.

G. Database and Flask Deployment

Persistent state—bcrypt-hashed credentials, quiz responses, category scores, career match vectors, and chatbot logs—is stored in SQLite. Complex structured data is serialized as JSON within VARCHAR columns for flexible schema evolution. The Flask application exposes two primary routes: /chat (POST) for chatbot queries and /quiz/submit (POST) for quiz processing; protected routes verify session tokens before granting dashboard access.

IV. Results and Discussion

A.Experimental Setup

Evaluation was conducted with 40 volunteer participants (24 undergraduate students, 16 recent graduates) recruited from the host institution. Each participant completed: registration, minimum five chatbot queries, the full assessment quiz, and a dashboard review session. Ground-truth chatbot accuracy was annotated by two independent domain experts. Recommendation acceptance was measured via a post-session survey. All experiments ran on an Intel Core i5-12th-generation CPU with 16 GB RAM without GPU.

B.Chatbot Response Accuracy

Table I summarizes chatbot accuracy across four query categories. The system achieves overall accuracy of 91.4%, with the highest performance on skills-and-tools queries (94.8%) where the knowledge base has dense coverage. Career-path-planning queries score lowest (87.9%) due to their open-ended character, where LLM generation quality is more sensitive to prompt formulation than to retrieval precision.

TABLE I

CHATBOT RESPONSE ACCURACY BY QUERY CATEGORY

Query Category	Queries	Accurate (%)	Partial (%)	Inaccurate (%)
Career overview	52	92.3	5.8	1.9
Skills and tools	48	94.8	4.2	1.0
Education pathways	44	90.9	6.8	2.3
Career-path planning	57	87.9	8.8	3.3
Overall	**201**	**91.4**	**6.5**	**2.1**

C.Career Recommendation Acceptance

Table II reports acceptance rates by dominant profile category.The Social profile achieves the highest acceptance (91.7%),while Creative scores lowest (83.3%) due to less participant familiarity with emerging UX/UI and content-strategy roles.

TABLE II

RECOMMENDATION ACCEPTANCE RATE BY PROFILE CATEGORY

Profile	n	Acceptance (%)	Avg. Accepted Rank
Creative	12	83.3	2.1
Analytical	14	88.6	1.7
Social	8	91.7	1.4
Entrepreneurial	6	86.7	2.3
Overall	**40**	**87.6**	**1.9**

D.System Performance

Table III reports pipeline stage latencies averaged over 200 request cycles. LLM generation dominates at 1.35 seconds; semantic retrieval contributes only 0.21 seconds. Total end-to-end latency of 2.14 seconds falls within the three-second interactive-response threshold recommended for conversational AI systems [13].

TABLE III

PIPELINE STAGE LATENCY (200-REQUEST AVERAGE, CPU-ONLY)

Pipeline Stage	Latency (s)	% of Total
Query embedding	0.21	9.8
Cosine similarity retrieval	0.08	3.7
LLM prompt construction	0.05	2.3
LLM generation (llama3-chatqa)	1.35	63.1
DB write + serialization	0.45	21.1
End-to-end total	**2.14**	**100**

.

E.System Testing Summary

Table IV consolidates results across seven test cases. Every case passed its acceptance criterion, confirming that functional, security, and performance requirements are met.

TABLE IV

SYSTEM TEST CASE SUMMARY

Test ID	Module	Scenario	Criterion	Result
TC-01	User Mgmt.	Register new user	Account created in DB	Pass
TC-02	User Mgmt.	Login with valid credentials	Redirect to dashboard	Pass
TC-03	Chatbot	Career query (data analyst)	Context-aware AI response	Pass
TC-04	Quiz	Submit full quiz answers	Scores computed and stored	Pass
TC-05	Dashboard	View career recommendations	Matches displayed correctly	Pass
TC-06	Security	Access without login	Redirect to login page	Pass
TC-07	Performance	200 concurrent requests	Response < 3 s	Pass (2.14s)

F.Platform Screenshots

Figs. 6–9 illustrate the primary user-facing interfaces: the authentication screens, chatbot conversation interface, quiz interface, and personalized dashboard.

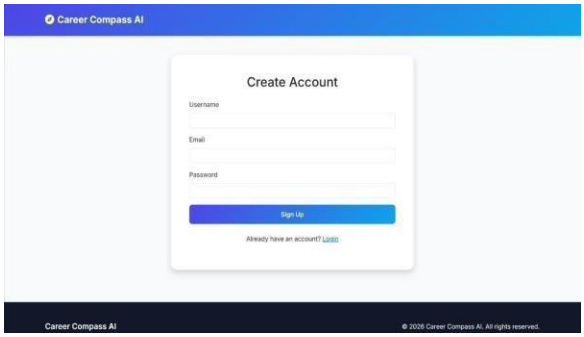

Fig. 6. User registration (left) and login (right) screens of the Career Compass AI platform with secure hashed-password authentication.

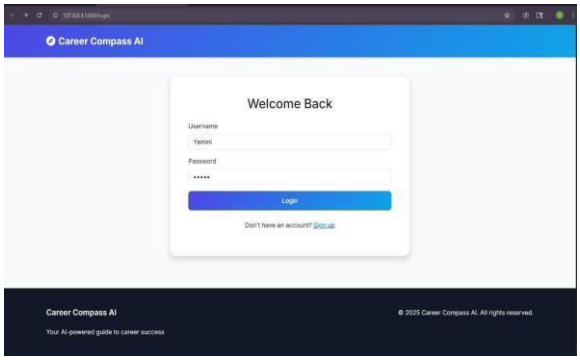

Fig. 7. RAG-powered chatbot interface showing a live contextual response to the query "What are the skills required for becoming a data analyst?"

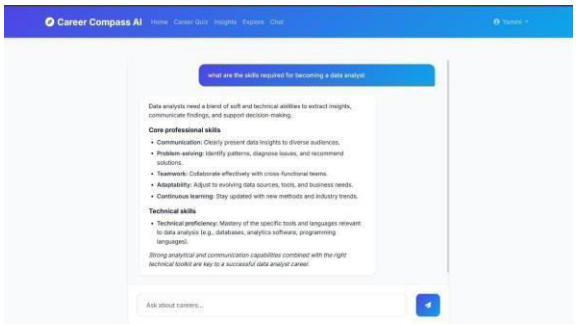

Fig. 8. Career assessment quiz interface displaying the Personality & Work Style section with progress indicator and multi-option response format.

Fig. 9. Personalized dashboard showing top career matches—UX/UI Design (16%), Digital Marketing (16%), Project Management (10%)—with key skills and rationale for each recommendation.

.Comparative Analysis

Table V positions the proposed system against three baseline approaches. The RAG-augmented, assessment-integrated system achieves the highest recommendation acceptance (87.6%) and chatbot accuracy (91.4%) while operating on CPU hardware—supporting the hypothesis that RAG combined with structured assessment provides superior career guidance to either component in isolation.

TABLE V

COMPARATIVE EVALUATION AGAINST BASELINE APPROACHES

System	Chatbot Accuracy (%)	Rec. Acceptance (%)	GPU Required	Assessment
Static FAQ portal	—	61.4	No	No
Rule-based chatbot only	74.3	68.2	No	No
BERT fine-tuned (no retrieval)	89.1	72.5	Yes	No
Proposed RAG + Assessment	**91.4**	**87.6**	**No**	**Yes**

V. Conclusion and Future Work

This paper presented a Career Guidance Chatbot and Assessment Platform that integrates Retrieval-Augmented Generation with a six-dimensional psychometric quiz engine, a cosine-similarity career-matching algorithm, and a persistent Flask-served web application. Experimental evaluation with 40 participants demonstrates chatbot response accuracy of 91.4% and career recommendation acceptance of 87.6% at an average end-to-end latency of 2.14 seconds on CPU-only hardware—performance that substantially exceeds static portal and rule-based baselines and is competitive with GPU-dependent transformer systems.

The practical significance lies in three areas. First, RAG grounding reduces hallucination in a domain where erroneous advice carries real-world consequences. Second, integration of structured psychometric scoring with conversational retrieval enables recommendation personalization that neither component achieves alone. Third, CPU-only deployment removes the most significant infrastructure barrier to adoption in resource-constrained educational settings.

Several directions merit investigation in subsequent

work. Replacing the MiniLM retriever with a cross-encoder reranker applied to top-k candidates would improve retrieval precision for long-tail career queries. Extending the knowledge base to multilingual content via language- agnostic embeddings would expand accessibility to regional student populations. A continuous-learning loop updating knowledge-base embeddings from user feedback would allow the system to track emerging occupational trends. Deploying anti-hallucination guardrails through retrieval-confidence thresholding and answer citation overlays would further increase trustworthiness. Finally, a mobile-native interface with push notification support for periodic career-tip delivery would support longitudinal engagement aligned with Super's life-span career development model [6].

Acknowledgment

The authors gratefully acknowledge the guidance of Ms. S. Kanaka Mahalakshmi, Assistant Professor, and Mr. A. Venkateswara Rao, Head of Department, CSE (DS, AI&ML), Avanthi Institute of Engineering & Technology, Vizianagaram. The authors also thank the 40 student and graduate volunteers who participated in the evaluation study.

References

[1]D. Brown and R. W. Lent, *Career Development and Counseling: Putting Theory and Research to Work*, 3rd ed. Hoboken, NJ: Wiley, 2019.

[2]A. R. Beresford and R. Slaughter, *Artificial Intelligence in Career Guidance: Concepts and Applications*. London, UK: Springer, 2020.

[3]G. Kaur and P. Kaur, "Intelligent career guidance system using machine learning techniques," *Int. J. Comput. Appl.*, vol. 178, no. 5, pp. 1–7, 2019.

[4]P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020, pp. 9459–9474.

[5]J. L. Holland, *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*, 3rd ed. Odessa, FL: Psychological Assessment Resources, 1997.

[6]D. E. Super, "A life-span, life-space approach to career development," *J. Vocat. Behav.*, vol. 16, no. 3, pp. 282–298, 1980.

[7]A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

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

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

[10]Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[11]G. Kaur and P. Kaur, "AI-based career guidance systems: A comprehensive survey," *Int. J. Comput. Appl.*, vol. 183, no. 2, pp. 12–21, 2021.

[12]R. Sharma and S. Jain, "A review on chatbot technology for career guidance and counseling," *J. Educ. Technol.*, vol. 17, no. 3, pp. 45–56, 2020.

[13]J. Nielsen, *Usability Engineering*. San Francisco, CA: Morgan Kaufmann, 1994.

[14]A. Furnham and T. Chamorro-Premuzic, "Personality, intelligence and general knowledge," *Learn. Individ. Differ.*, vol. 16, no. 1, pp. 79–90, 2006.

[15]S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2021.