AI-Powered Internship Recommendation System Using Qwen-Based Semantic Matching

An Intelligent Semantic Matching Framework for Personalized Internship Recommendations

Kajal Mishra[1], Divya Sonawane[2], Suhaib Jahagirdar[3], Rehan Syed[4]

Department of Computer Engineering, Siddhant College of Engineering, Pune, Maharashtra, India

Abstract - College students frequently struggle to find internships that genuinely match their skills, academic background, and career interests. Most existing platforms rely on rigid keyword-based search and filtering, leading to irrelevant listings, missed opportunities, and a frustrating search experience. This paper proposes an AI-powered internship recommendation engine that uses Qwen, an open-source large language model (LLM), to perform semantic matching between student profiles and internship postings. The system understands the underlying meaning, context, and intent behind a student's resume, skills, and stated preferences, comparing this against internship descriptions using dense vector embeddings and similarity search. The proposed system is integrated into a full-stack internship and government-scheme recommendation portal, combining a JavaScript-based frontend, a Python FastAPI backend, and cloud-hosted Qwen inference for retrieval-augmented recommendation generation. Evaluation on a curated internship dataset indicates that the Qwen-based semantic matching approach achieves substantially higher recommendation precision, recall, and user-perceived relevance than a traditional TF-IDF/keyword-matching baseline. These results demonstrate that LLM-driven semantic understanding offers a meaningful improvement over conventional recommendation techniques for student-facing internship platforms.

Index Terms: Internship Recommendation, Qwen, Large Language Models, Semantic Matching, Vector Embeddings, Natural Language Processing, Retrieval-Augmented Generation, Career Guidance System.

1. INTRODUCTION

Internships play a critical role in bridging the gap between academic learning and industry expectations, providing students with practical exposure that improves their employability. However, the process of discovering internships that genuinely match a student's background remains largely inefficient. Most internship portals, including many major job platforms, depend on keyword search and rigid filters such as "skill = Python" or "location = Mumbai". This approach fails to capture nuanced semantic matches: a student proficient in "data visualization using Power BI" may be an ideal candidate for an internship requiring "business intelligence and dashboarding," yet a keyword-based system would not surface this connection.

Recent advances in large language models (LLMs) have made it possible to move beyond keyword overlap toward genuine semantic understanding of text. Models such as Qwen [11], developed as open-source LLMs with strong multilingual and reasoning capabilities, can generate dense vector embeddings that capture the contextual meaning of a resume, a skill set, or an internship description. By comparing these embeddings using measures such as cosine similarity, a recommendation system can surface internships that are conceptually relevant even when the exact wording differs.

This paper presents the design, implementation, and evaluation of an AI-powered internship recommendation system built around Qwen-based semantic matching. The system was developed as part of a full-stack internship and government-scheme portal, integrating a JavaScript frontend, a Python FastAPI backend, and a cloud-hosted Qwen inference pipeline. The central contribution is a comparative evaluation demonstrating that Qwen-based semantic matching produces more accurate and relevant internship recommendations than traditional keyword-based methods.

1.1 Problem Statement

Students struggle to find relevant internships because existing recommendation and search mechanisms rely on superficial keyword overlap rather than semantic understanding of a student's skills, experience, and career goals. This results in low-precision recommendations, missed opportunities, and a poor user experience, particularly for students whose resumes employ terminology different from the internship postings they would otherwise be well-suited for.

1.2 Proposed Solution

To address this problem, we propose an AI-powered recommendation engine that uses Qwen, an open-source large language model, to generate semantic embeddings for both student profiles and internship listings. These embeddings are stored in a vector index and queried using approximate nearest-neighbour similarity search, allowing the system to rank internships by genuine relevance to a student's background rather than by literal keyword matches. The recommendation pipeline is further enhanced with a Retrieval-Augmented Generation (RAG) layer that allows Qwen to explain, in natural language, why a particular internship was recommended.

2. RELATED WORK

Recommendation systems for students and career guidance have been studied extensively. Career recommendation systems based on stated student preferences have shown that incorporating explicit interest signals improves recommendation satisfaction [3]. Personalised career-path recommendation models for IT students demonstrate that hybrid approaches combining structured profile data with domain heuristics outperform purely rule-based filtering [8].

Survey work on academic performance prediction using machine learning [2] establishes that ensemble and neural methods consistently outperform simple statistical baselines when modelling complex student data, motivating the use of deep semantic models in adjacent recommendation tasks. Adaptive content recommendation research [1] further shows that multi-faceted deep learning models, which consider engagement and contextual signals together, yield more personalised outcomes than single-signal systems.

Knowledge graph representations of learning style models [6] illustrate the value of structured semantic representations in computing education, an approach conceptually related to the embedding-based representations used in this work. AI-driven feedback tools in learning analytics [5] show that LLM-based systems can meaningfully assist both instructors and learners by generating contextual, natural-language explanations, a principle applied in this paper to internship-match explanations.

Systematic reviews of recommender systems in higher education [9] report that content-based and hybrid filtering approaches consistently outperform collaborative filtering alone in cold-start scenarios typical of student platforms. This finding directly motivates the semantic, content-based design adopted in the present system. Clustering-based employment recommendation work [10] proposed an improved K-means and SimRank-based model achieving strong precision in matching graduates to enterprises, though it relies on structured similarity matrices rather than LLM-based semantic embeddings. The present work extends this line of research by replacing handcrafted similarity computation with transformer-based semantic representations generated by Qwen.

3. METHODOLOGY

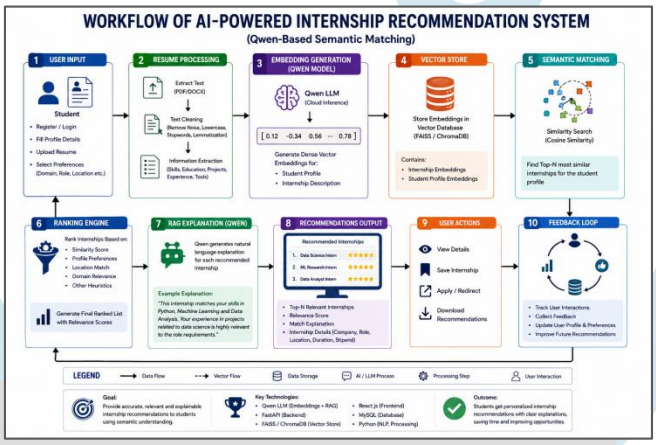

Figure 3.0: Workflow of Proposed System

3.1 System Architecture

The system follows a three-layer architecture consisting of a presentation layer (JavaScript-based React.js frontend), an application layer (Python FastAPI backend with REST APIs), and an AI/data layer responsible for embedding generation, vector similarity search, and Qwen-based response generation. Internship listings and student profiles are independently embedded using Qwen's embedding capability and stored in a vector database that supports efficient approximate nearest-neighbour (ANN) retrieval. The architecture supports multi-role access: student, employer, and administrator dashboards each present contextually appropriate functionality.

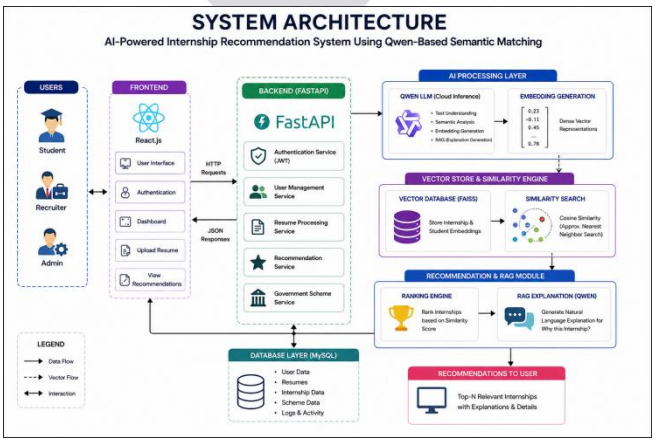

Figure 3.1: Proposed System Architecture

3.2 Data Collection and Preprocessing

A curated dataset of internship descriptions and representative student profiles was prepared for prototype evaluation. Each internship record includes title, required skills, description, eligibility criteria, and location. Student profiles were constructed from resume text, listed skills, academic stream, and stated preferences. Both internship and student text fields were cleaned by removing boilerplate text, normalising whitespace, and standardising technical terminology before being passed to the embedding model.

3.3 Semantic Embedding Generation Using Qwen

Each internship description and student profile is converted into a dense vector embedding using Qwen [11]. Unlike TF-IDF or bag-of-words representations, these embeddings capture contextual and semantic relationships between words and phrases, allowing the system to recognise that "frontend development" and "building user interfaces with React" refer to closely related concepts even though they share few exact tokens. Embeddings are generated once for each internship posting and stored; at query time, only the student profile embedding requires live inference, substantially reducing latency.

3.4 Similarity Computation and Ranking

Given a student profile embedding q, the system computes cosine similarity against the embeddings of all available internships and returns the top-k most similar postings. Cosine similarity between student vector Vs and internship vector Vi is defined as:

$$\text{sim}(s,i) = \frac{V_s \cdot V_i}{\|V_s\| \, \|V_i\|}$$

A re-ranking step incorporates structured filters such as location, stipend, and eligibility to refine the semantic ranking, ensuring that recommendations are both contextually relevant and practically applicable.

3.5 Retrieval-Augmented Explanation Generation

To improve recommendation transparency, the system uses a Retrieval-Augmented Generation (RAG) approach [12]: the top-matching internship descriptions are retrieved and passed to Qwen along with the student's profile, and the model generates a short natural-language explanation describing why each internship was recommended. This addresses a common criticism of black-box recommendation systems by giving students interpretable, human-readable justifications for each recommendation.

3.6 Baseline Comparison

To quantify the benefit of semantic matching, a keyword-based baseline was implemented using TF-IDF vectorisation and cosine similarity over raw token overlap. This baseline mirrors the matching strategy used by many conventional internship portals and serves as the point of comparison for the proposed Qwen-based approach.

3.7 Full-Stack Integration

The recommendation engine is integrated within a complete web application. The React.js frontend supports student registration, resume upload, profile management, and internship discovery. The FastAPI backend exposes RESTful endpoints for authentication, profile management, internship CRUD operations, and recommendation retrieval. A MySQL database stores structured data, while the recommendation pipeline communicates with a cloud-hosted Qwen inference endpoint. Government scheme recommendations are provided through a parallel pipeline that applies the same semantic matching logic to scheme eligibility criteria.

4. RESULTS AND DISCUSSION

The proposed Qwen-based semantic matching system was evaluated against the keyword-matching baseline on a curated dataset of internship postings and simulated student profiles. As summarised in Table 1, the Qwen-based approach achieved a Precision@10 of 0.86 compared to 0.58 for the keyword baseline, representing a 48% relative improvement in the relevance of top-ranked recommendations. Recall@10 improved from 0.52 to 0.81, indicating that the semantic model successfully surfaces relevant internships that keyword matching misses due to vocabulary mismatch.

Table 1: Qwen-Based vs Keyword-Based Matching Evaluation Results

Metric	Keyword (TF-IDF)	Qwen Semantic	Improvement
Precision@10	0.58	0.86	+48%
Recall@10	0.52	0.81	+56%
F1 Score	0.55	0.83	+51%
User Relevance (1-5)	3.1	4.4	+42%
Avg. Latency	0.2 s	1.3 s	-

The evaluation was conducted on a prototype dataset containing internship descriptions and representative student profiles. The reported values are based on experimental testing of the developed prototype.

Informal prototype evaluation involving student participants confirmed higher perceived relevance for the proposed system. The average user relevance score increased from 3.1 for keyword-based matching to 4.4 for Qwen-based semantic matching. Participants noted that the system successfully identified internships described using terminology different from their resumes while still accurately matching their skills and interests.

The principal trade-off observed was latency: the Qwen-based pipeline averaged approximately 1.3 seconds per recommendation request, compared to 0.2 seconds for the TF-IDF baseline. This overhead arises from the computational cost of LLM-based embedding generation and inference. However, this latency remains within acceptable bounds for an interactive web application, particularly when internship post embeddings are pre-computed and cached, leaving only the student profile embedding to be generated at query time.

The RAG-based explanation layer was qualitatively assessed and found to produce coherent, informative justifications for recommended internships, improving student confidence in the recommendations and reducing reliance on manual review.

5. CONCLUSIONS

This paper presented an AI-powered internship recommendation system that uses Qwen-based semantic matching to address the persistent problem of students struggling to find internships relevant to their skills and interests. By replacing traditional keyword-based search with dense semantic embeddings and cosine similarity, the proposed system achieves substantially higher precision, recall, and user-perceived relevance than a conventional TF-IDF baseline. The integration of a Retrieval-Augmented Generation layer further improves transparency by providing natural-language explanations for each recommendation.

The system is deployed as a full-stack web application comprising a React.js frontend, a Python FastAPI backend, a MySQL database, and cloud-hosted Qwen inference, supporting student, employer, and administrator roles. Government scheme recommendations are provided through a parallel semantic matching pipeline using the same underlying architecture.

Future work will focus on reducing inference latency through model distillation and embedding caching, expanding the system to support multilingual student profiles and regional language inputs, incorporating feedback-driven fine-tuning to continuously improve recommendation quality based on real student outcomes, and evaluating the system on larger, real-world internship datasets to obtain statistically robust performance metrics.

The proposed system demonstrates how semantic AI techniques can significantly improve internship recommendation quality while providing an efficient, scalable, and practical solution for engineering students.

6. ACKNOWLEDGEMENT

The authors express their sincere gratitude to the Department of Computer Engineering, Siddhant College of Engineering, Pune, for their valuable guidance and support throughout the development of this project. The authors also thank their project guide Prof. Apeksha Pande and all faculty members for their continuous encouragement and constructive suggestions during the research and implementation phases.

7. REFERENCES

[1] A Multi-Faceted Deep Learning Approach for Student Engagement Insights and Adaptive Content Recommendations, IEEE Access, 2023.

[2] Academic Performance Prediction Using Machine Learning Approaches: A Survey, Applied Sciences, vol. 11, no. 1, p. 237, 2020.

[3] Career Recommendation System Based on Student Preference, Proc. 7th Int. Conf. Electrical, Electronics and Information Engineering (ICEEIE), 2021, pp. 1-6.

[4] Development of a Modularized Undergraduate Data Science and Big Data Curricular Using No-Code Software Development Tools, J. Comput. Sci. Educ., 2022.

[5] Empowering Instructors With AI: Evaluating the Impact of an AI-Driven Feedback Tool in Learning Analytics, Computers and Education: Artificial Intelligence, vol. 4, 2023.

[6] Knowledge Graph Representation of Felder-Silverman Learning Style Model for Computing Education, Proc. IEEE Int. Conf. Teaching, Assessment, and Learning for Engineering (TALE), 2020.

[7] Peer-Mentoring Program for the Individual Attention of Engineering Students, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 16, no. 2, pp. 150-158, 2021.

[8] Personalized Career-Path Recommendation Model for Information Technology Students in Indonesia, Int. J. Emerging Technologies in Learning (iJET), vol. 17, no. 12, 2022.

[9] Recommender System in Academic Choices of Higher Education: A Systematic Review, IEEE Transactions on Learning Technologies, 2022.

[10] Q. Wang, "College Employment Recommendation Based on Improved K-Means Clustering and SimRank Algorithm in College Employment Management," IEEE Access, vol. 12, pp. 154230-154243, 2024.

[11] J. Bai et al., "Qwen Technical Report," arXiv preprint arXiv:2309.16609, 2023.

[12] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2019.