Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

RAG-Enhanced LLM Job Recommendation Systems: Balancing Efficiency and Accuracy in Candidate–Job Matching

Vaibhav Arora, Dr. Sunil Maggu
Maharaja Agrasen Institute of Technology, Delhi

Abstract **- To improve candidate–job matching, this research develops a job recommendation system called JobPilot, which uses vector-based semantic retrieval with Large Language Models (LLMs). JobPilot's architectural design takes a simpler yet more effective CRUD approach, using text embeddings and a vector database (Pinecone) to retrieve semantically similar jobs, rather than overly complex production pipelines that rely on message queues or multi-stage orchestration. The Gemini LLM provides the final ranking and explanation, while a pre-trained transformer model (Xenova/all-MiniLM-L6-v2) generates the embeddings. The dataset contains 10 user profiles and 1,000 job postings. Although empirical accuracy metrics are not calculated, the system design is based on recent studies showing the benefits of embedding-based retrieval and LLM-based reasoning in recommender settings. The paper discusses implementation details, theoretical benefits, limitations and future directions.**

Keywords: Job Recommendation, LLM, RAG, Embeddings, Pinecone, CRUD System, AI Matching

1. INTRODUCTION -

In digital recruitment platforms, job-recommendation systems are a key mechanism for facilitating the connection of employers with candidates. Traditional systems often use collaborative filtering or keyword matching. The drawback of these methods is that they struggle with semantic nuances—for example, "software engineer" might match with "backend developer." Transformer-based embeddings and LLMs have enabled much better and more understandable matching through deeper contextual relevance in resumes and job descriptions. This paper presents JobPilot, a job-recommendation system based on CRUD methodology that 're-queues' based on LLM and semantic embeddings, and the goal is to demonstrate how an RAG-style pipeline, separate from a question-inferential evaluation of recommendation, can improve the output of recommending a candidate without needing to establish significant infrastructure. The pipelines we utilise are focused on semantic retrieval, not evaluation, using a dataset of 1,000 job postings and 10 candidate profiles.

2. LITERATURE REVIEW

2.1 RAG and Recommendation Systems

Retrieval-Augmented Generation allows for improvement in fact grounding and contextual understanding of the LLMs due to retrieval before generation in a direct sense [1]. In the recommendation systems, RAG helped ensure that the generated explanation rankings were actually based on actual documents retrieved and not from model memory. [2].

2.2 Embedding-Based Retrieval

Sentence-transformer-based models, such as MiniLM and S-BERT, have outperformed keyword-based methods in semantic similarity-related tasks in several studies [3][4]. Their effectiveness is also demonstrated in resume–job matching studies where embeddings capture context beyond surface terms [5]. Recent works like Resume2Vec show that embedding-driven systems enhance the relevance between candidates and jobs substantially [6].

2.3 Vector Databases and Approximate Nearest Neighbours

Vector databases like Pinecone support high-speed semantic search using approximate nearest-neighbour (ANN) algorithms [7]. Such systems allow scalability while maintaining low latency. The trade-off between recall, latency, and index type is key for practical deployments [8].

2.4 LLMs for Explainability and Ranking

LLMs have become interpretation devices for recommendation systems. For instance, studies have shown that LLMs provide natural language explanations for recommendations by improving user trust in them [9]. Furthermore, it has come to light that grounding these

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

explanations using RAG reduces the hallucination and makes the produced text more reliable [10].

2.5 Resume Parsing and Representation

This requires meaningful embeddings, and accurate parsing of the resume is necessary for that. The transformer-based parsers have extracted skills and experiences while reducing keyword overloading problems. By combining structured and unstructured features, candidate profiling can be done robustly [6] [9].

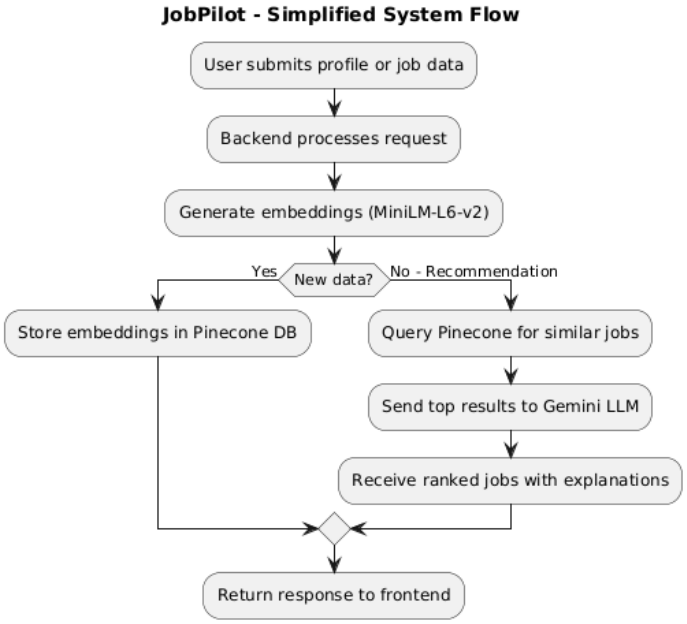
Therefore, combining embeddings, vector retrieval, and the LLM reasoning together may be considered most effective for effective and explainable recommendations.

3. SYSTEM DESIGN AND IMPLEMENTATION

3.1 Overview

JobPilot implements this using a CRUD architecture with a Node.js/Express backend and Pinecone as the vector database. The goal is simplicity while availing modern NLP techniques for semantic job matching.

System Workflow:



3.2 Data Description

The dataset includes:

- More than 1,000 job postings with structured fields: title, company, location, required skills, and description.
- 10 candidate profiles, each with parsed skills and summarised experiences.

The data is stored and managed by default CRUD operations: add, read, update, and delete.

3.3 Embedding and Retrieval Process

Both the job descriptions and the resumes are embedded using Xenova/all-MiniLM-L6-v2 and then generate 384-dimensional vectors; these vectors are indexed in Pinecone, using cosine similarity as the distance metric.

For each candidate:

- Similarity search by Pinecone pulls the top 10 job vectors.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

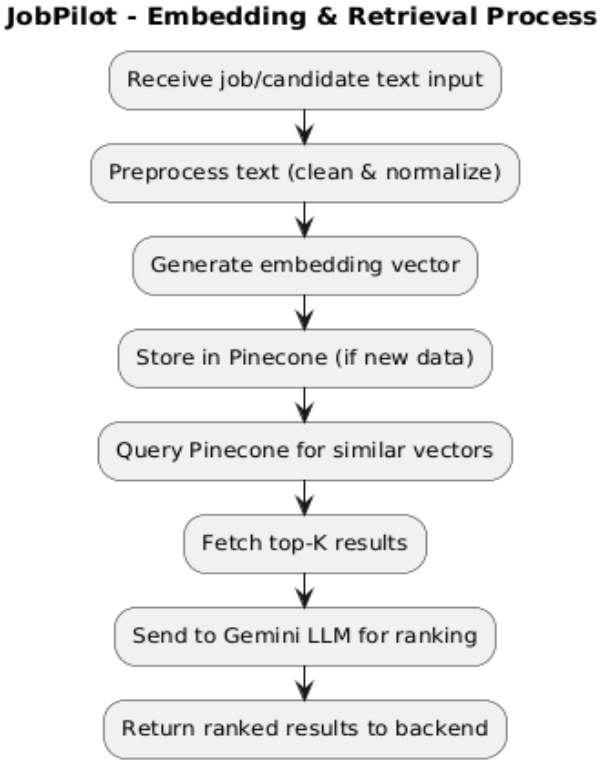
- Results retrieved are passed to Gemini LLM for ranking and the generation of explanations for why each job fits the profile.

3.4 LLM-based Re-Ranking

The Gemini LLM is given a retrieved job detail with the context of a candidate's resume. It outputs :

- A relevance score (0–1)
- A short explanation, such as "This job matches because your Python and backend experience align with the requirements."

3.5 Pipeline:



4. DISCUSSION

The JobPilot system illustrates how an RAG-inspired architecture, in turn highly simplified, can make effective use of modern NLP tools. Lacking message queues or caching layers for simplicity, it still captures the semantic relationships between candidates and jobs.

Advantages:

- Semantic understanding beyond keyword matching
- Interpretable recommendations by LLM explanations
- Easy to deploy (CRUD architecture)

Limitations:

- Small dataset-no formal accuracy metrics, such as Recall@K, NDCG@K
- LLM responses are diverse and non-deterministic
- Embedding updates needed for dynamic job databases

5. CONCLUSION AND FUTURE WORK

JobPilot demonstrates a practical and understandable way to integrate embeddings and LLMs for job recommendations on a small-scale dataset. The CRUD-based structure makes it

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

accessible for experimentation without complex infrastructure.

Future directions include:

- Adding larger, labelled datasets to which error can be measured
- Incorporating fairness and bias analysis
- Explore hybrid embeddings, such as late-interaction methods, for finer rankings.
- Optimising LLM prompting for consistency as well as reliability

REFERENCES

[1] Wei, W., Duan, H., Zhuo, X., Wang, K., Huang, Y., & Liu, X. (2025). Enhanced recommendation systems with retrieval-augmented large language model (ER2ALM). Journal of Artificial Intelligence Research.
[2] Wang, S., Fan, W., Feng, Y., Ma, X., Wang, S., & Yin, D. (2025). Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation (K-RagRec). arXiv preprint.
[3] Biswas, P., Hall, A., & Moll, F. (2024). Enhanced resume screening for smart hiring using Sentence-BERT (S-BERT). International Journal of Advanced Computer Science and Applications, 15(8).
[4] Zhao, X., Wang, M., Li, J., Zhou, S., Yin, D., Li, Q., Tang, J., & Guo, R. (2023). Embedding in recommender systems: A survey. arXiv preprint.
[5] Ma, Q., Ren, X., & Huang, C. (2024). XRec: Large language models for explainable recommendation. Findings of EMNLP 2024.
[6] Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., Lund, B., Xiao, T., Pasem, H., & Rupeshkumar, S. (2025). Resume2Vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching. Electronics, 14(4).
[7] Ma, L., Zhang, R., Han, Y., Yu, S., Wang, Z., Ning, Z., & Zhou, Y. (2023). A comprehensive survey on vector databases: Storage and retrieval techniques. IEEE Journal on Big Data.
[8] Santhanam, K., Bai, H., & Efros, A. A. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. NAACL 2022.
[9] Peng, Q., Liu, H., Huang, H., Yang, Q., & Shao, M. (2025). A survey on LLM-powered agents for recommender systems. arXiv preprint.
[10] Yu, X., Xu, R., Xue, C., Zhang, J., Ma, X., & Yu, Z. (2025). CONFIT V2: Improving resume-job matching using hypothetical resume embedding and runner-up hard-negative mining. Findings of ACL 2025.