



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Literature Survey on AI Job Recommendation System Using LLM and MCP Server

Premashree R
PG Scholar (M.Tech)
Dept of Computer Science and Engineering.
SJC Institute of Technology.

Manjunatha Kumar B H
Professor and Head of Department
Dept of Computer Science and Engineering
SJC Institute of Technology.

Abstract

The rising popularity of Artificial Intelligence has caused a paradigm shift in job recommendation systems, making it possible to have more precise and personalized recommendations. Conventional recommendation systems based on traditional Machine Learning and Deep Learning architectures have been hampered due to limitations in capturing semantic meaning while understanding candidate resumes and job descriptions. However, recent breakthroughs and developments based on Large Language Models (LLMs) have tackled these issues with better natural language understanding and reasoning. The research aims to include a comprehensive literature review on emerging advancements related to LLM-based job recommendation systems, including advancements based on semantic profiling and rankings, explainable recommendation accuracy, and LLM-agents-based workflows. It also investigates MCP server role-based emerging frameworks as an integrative platform for LLMs with secured access toward organizational tools, databases, and job services. According to recent research reports and trials, MCP LLM Networks have promoted better system scalabilities and temporal freshness with better dependability and trustworthiness compared with previous LLM-based isolated pipeline architectures. Despite these benefits, research and advancements are still necessary within LLMs with regard to challenges associated with fairness and precision, computation costs, and common testing metrics.

Key Words: Large Language Models (LLMs), Model Context Protocol (MCP) Server, Recommender Systems, Real-Time Job Matching.

Introduction

The rapid development of AI techniques has transformed the process by which organizations match job seekers with appropriate employment opportunities. Conventional systems for recommending jobs relied on keyword matching, rule-based filtering, or algorithms from classical machine learning-typically without much

understanding of the deep semantics of resumes and job descriptions. With the advent of Large Language Models (LLMs), such systems have now started to understand textual context, extract skills with better accuracy, and present highly personalized suggestions. This paper presents a literature review that surveys the current state of LLM-based integration in job recommendation architectures and discusses how these new generation models enhance overall efficiency and intelligence in candidate-job matching.

The relevance and usefulness of LLM-based job recommendation systems have significantly escalated because of the rising number of job postings and varying candidate profiles. Employers and job seekers are also demanding higher relevance and capabilities from job recommendation systems that offer insights, recommendations, and updates. LLMs possess impressive capabilities in understanding languages, making it possible for them to process market trends and generate insights and recommendations. Because businesses are increasingly adopting automation and intelligent recruitment processes, adoption of LLMs becomes integral.

In these years, MCP has evolved in recent years to become an important way through which LLMs are securely connected with external tools, databases, and enterprise systems. While the LLM offers strong reasoning and understanding, it needs to get real-time contextual information to provide job recommendations that are accurate. MCP servers bridge this gap by allowing LLMs to securely access ATS, skill taxonomies, job repositories, and organizational datasets-all without retraining the model. All this seamless integration promises to make job recommendation systems dynamic, reliable, and operationally scalable with enormous improvements compared to traditional AI pipelines.

The literature survey scope will be centred on examining the existing research on LLM-driven recommendation systems, job recommendation methods, parsing procedures for resume and job description data, and generation models with retrieval augmentation. It will also incorporate recent research and technical reports on MCP servers and their utilization in improving context-driven intelligent systems. The literature will be able to identify

some of the emerging trends, benefits, and limitations associated with existing research and technologies, and more specifically on the utilization of LLMs and MCP infrastructure for developing next-generation job recommendation systems. By examining academic research and latest industry advancements, a well-rounded understanding of current state-of-the-art research will be achieved.

1. Background and Key Concepts

A Job Recommendation System (JRS), with its end goal being the mapping of qualified candidates with appropriate job opportunities, requires the simultaneous modelling of candidate profiles and job postings. Large Language Models (LLMs), on the other hand, are transformer neural networks pre-trained on big text datasets and employed within our framework for semantic reasoning, information extraction, and natural-language processing. Model Context Protocol (MCP), an open-protocol framework, facilitates communication between LLM hosting services and external “MCP servers” or tools/data sources. These allow LLMs to effectively obtain real-time context and services and integrate tool-supported workflows without requiring custom coding. These three concepts, JRS, LLMs, and MCP servers, conceptualize modern recommendation systems.

Two complementary theoretical frameworks have become dominant when reasoning about LLM-enabled recommender systems. One distinguishes discriminative and generative paradigms: On the one hand, discriminative approaches treat LLMs as powerful encoders or scorers that output embeddings or probability scores for candidate–job pairs, useful for ranking and classification; on the other hand, generative approaches use LLMs to generate ranked recommendations, explanations, or multi-step plans in natural language that may be further parsed into actions. This taxonomy helps explain trade-offs: discriminative methods are often cheaper to deploy as encoders, whereas generative methods provide richer explanations and flexible dialog but at higher cost and risk of hallucination. Recent surveys formalize this DLLM4Rec vs. GLLM4Rec distinction and present many hybrid designs which combine encoder-based retrieval by generative re-ranking.

A. AI-Based Job Recommendation Systems

Job recommendation systems aim to assist job seekers by suggesting relevant job opportunities based on their skills, experience, preferences, and career goals. Early systems relied on information retrieval techniques such as keyword matching and Boolean queries. With the growth of data availability, machine learning models such as collaborative filtering and content-based filtering were adopted to improve recommendation quality. However, these approaches often struggle with sparse data, cold-start problems, and limited semantic understanding of textual content. Recent research has therefore shifted toward deep learning and natural language processing techniques to better analyse resumes and job descriptions.

Large Language Models represent a major advancement in this domain by enabling deep semantic understanding and contextual reasoning. LLMs can interpret unstructured text, infer implicit skills, and generate personalized explanations for recommendations. Studies published during 2024 and 2025 demonstrate that LLM-based systems outperform traditional recommender models in tasks such as skill extraction, job matching, and candidate

profiling. These models also enable conversational job recommendation, allowing users to interact with the system in natural language and receive tailored career guidance.

B. Large Language Models and Model Context Protocol (MCP)

Large Language Models are transformer-based architectures trained on massive text corpora, enabling them to perform tasks such as text understanding, summarization, and reasoning. In job recommendation systems, LLMs are commonly used for resume parsing, job description normalization, ranking, and explainable recommendations. However, a major limitation of LLMs is their inability to directly access real-time enterprise data or external systems. This limitation restricts their effectiveness in dynamic recruitment environments where job postings and candidate profiles change frequently.

The Model Context Protocol (MCP) addresses this limitation by providing a standardized interface for connecting LLMs to external services through MCP servers. MCP servers act as intermediaries that expose tools such as job databases, applicant tracking systems, skill taxonomies, and analytics engines to LLM-based applications. By enabling secure tool-calling and real-time context retrieval, MCP enhances the reliability, scalability, and practical deployment of LLM-powered job recommendation systems. Recent studies highlight MCP-enabled architectures as a promising direction for building intelligent, context-aware, and enterprise-ready AI recruitment platforms.

C. Research Motivation

The rapid expansion of online recruitment platforms and digital hiring processes has created an urgent need for intelligent job recommendation systems capable of handling large volumes of unstructured data. Traditional recommendation techniques, including rule-based and classical machine learning models, often fail to capture the semantic relationships between job requirements and candidate profiles. Although recent deep learning models have improved recommendation accuracy, they still lack the reasoning and contextual understanding required for personalized and explainable job suggestions. Large Language Models (LLMs) have emerged as a promising solution due to their ability to understand natural language, extract implicit skills, and generate context-aware recommendations. However, the practical adoption of LLMs in real-world recruitment systems remains limited due to challenges such as static knowledge, high computational cost, and limited access to real-time enterprise data.

The integration of Model Context Protocol (MCP) servers with LLM-based job recommendation systems provides a strong motivation for further research. MCP enables LLMs to securely interact with external databases, applicant tracking systems, and job repositories, allowing recommendations to be based on up-to-date and verified information. Despite this potential, existing literature lacks a unified analysis of how LLM reasoning capabilities and MCP-driven context access can be combined effectively. Issues related to scalability, fairness, privacy, and evaluation standards remain underexplored. This motivates the need for a comprehensive literature survey that consolidates recent research, identifies open challenges, and highlights future directions for building

reliable, context-aware, and intelligent AI job recommendation systems.

D. Existing Technologies

The earliest technologies used in AI job recommendation systems are based on information retrieval and traditional machine learning techniques. Keyword-based matching, Boolean search, and rule-driven filters were widely adopted in early recruitment platforms. These systems rely heavily on exact term matching between resumes and job descriptions, resulting in low semantic accuracy. To overcome these limitations, machine learning models such as logistic regression, decision trees, and collaborative filtering were introduced. While these models improved personalization to some extent, they struggled with sparse data, cold-start problems, and limited contextual understanding. As job descriptions and resumes are highly unstructured, these early technologies lacked the ability to infer implicit skills and career intent, motivating the shift toward deep learning approaches. Deep learning technologies marked a significant advancement in job recommendation systems by enabling semantic representation learning. Models such as Word2Vec, GloVe, and later BERT-based encoders were used to generate embeddings for resumes and job descriptions. These embeddings allowed systems to perform similarity-based matching using vector search techniques. Convolutional and recurrent neural networks were also explored for sequence modelling of career paths. Although these deep learning models significantly improved matching accuracy, they required extensive feature engineering and domain-specific tuning. Furthermore, they were typically limited to single-task objectives such as ranking, without offering interpretability or conversational interaction, which constrained user engagement. Large Language Models represent the current state-of-the-art technology in AI job recommendation systems. Transformer-based LLMs such as GPT-style models and instruction-tuned variants enable advanced natural language understanding, skill extraction, and reasoning across multiple documents. LLMs can parse resumes, normalize job titles, infer missing skills, and generate

personalized job recommendations in natural language. Recent systems employ prompt-based learning and few-shot techniques to adapt LLMs to recruitment tasks without extensive retraining. Additionally, LLMs support explainable recommendations by justifying why a particular job matches a candidate profile, thereby increasing transparency and user trust.

Retrieval-Augmented Generation (RAG) is another key technology used alongside LLMs in job recommendation systems. RAG architectures combine vector databases with LLMs to retrieve relevant job postings or candidate profiles before generating recommendations. This approach addresses the limitation of static knowledge in LLMs by grounding outputs in external documents. Vector databases such as FAISS and similar embedding-based retrieval systems are commonly used to index resumes and job descriptions. By integrating retrieval with generation, RAG-based job recommenders achieve higher factual accuracy, scalability, and relevance compared to standalone LLM approaches. This architecture is widely adopted in recent research and industry implementations.

Agent-based and tool-integrated LLM technologies are emerging as the next phase in AI job recommendation systems. In these systems, LLMs act as intelligent agents capable of invoking external tools, APIs, and enterprise services to complete recommendation tasks. Model Context Protocol (MCP) servers play a critical role by providing standardized and secure access to job databases, applicant tracking systems, and analytics tools. This enables real-time job matching, dynamic candidate profiling, and automated workflow execution. Such architectures enhance scalability, reduce integration complexity, and improve system reliability. Current research indicates that combining LLM agents with MCP-enabled infrastructures can significantly advance the deployment of intelligent, context-aware job recommendation platforms.

The following table provides a summary of merits and remarks of research works considered in this literature survey.

Paper Title & Year	Authors	Merits	Remarks
Generative Job Recommendations with LLMs (2023)	Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, Hui Xiong	Introduces a novel generative job recommender (GIRL) using LLMs with CV-to-JD generation; handles cold start & explainability.	Moves beyond traditional ranking to generate actionable job descriptions; more user-centric.
Large Language Models are Zero-Shot Rankers for Recommender Systems (2023)	Hou et al.	Demonstrates LLMs' potential as zero-shot ranking models for recommendation tasks.	Useful insight into using general LLM reasoning for ranking in job recommender pipelines.
Chat-REC: Interactive & Explainable LLM-Augmented Recommender (2023)	Gao, Sheng, Xiang et al.	Combines interactive LLMs with recommendation systems; enhances explainability & cross-domain capabilities.	Offers foundational ideas for conversational job recommenders.

RecRanker: Instruction-Tuned LLM as Ranker (2023)	Luo, He, Zhao et al.	Uses instruction fine-tuning to align LLM ranking with recommendation needs; hybrid ranking tasks.	Emphasizes training data quality and prompt design for reliable job ranking.
Large Language Models are Not Stable Recommender Systems (2023)	Ma, Cheng, Zhu, Xiong	Identifies instability and bias issues in LLM-based recommenders; proposes calibration framework.	Highlights key evaluation concerns for LLM-powered job recommenders.
LLM4Jobs: Unsupervised Occupation Extraction (2023)	Li, Kang, De Bie	Presents methodology for extracting and standardizing occupations with LLMs — crucial for matching stages.	Serves as a strong preprocessing step in job matching.
Enhance Large Language Models as Recommendation Systems (2025)	(Critic-LLM-RS)	Proposes web crawling + LLM integration for dynamic job recs.	Bridges classic RS principles with LLM generative strength.
Integrating LLMs with Near-Real-Time Data for Job Rec (2024)	(unnamed system)	Proposes web crawling + LLM integration for dynamic job recs.	Shows direction for real-time job data acquisition + semantic matching.
Balancing Efficiency & Accuracy in LLM Job Matching (2025)	(JobPilot paper)	LLM + semantic embedding + RAG pipeline for job matching	Combines embeddings and generative reasoning for relevance and explainability.
Enhancing E-Recruitment via Text Summarization & LLMs (2025)	(MDPI study)	Uses LLMs and summarization for job description processing.	Improves job/skill understanding in recommendation.
Job Recommender Systems: SLR & Challenges (2025)	Springer Big Data Review	Systematic review highlighting gaps up to 2023; suggests LLM future use.	Good reference baseline for job recommender evolution.
Recommender Systems Meet LLM Agents: Survey (2025)	Zhu et al.	Comprehensive survey of LLM in RS, including generative/connector paradigms.	Valuable for conceptual framing of LLM-driven recommendation workflows.
A Measurement Study of Model Context Protocol (2025)	Guo et al.	First empirical study of MCP ecosystem scale, risks, & server quality.	Provides MCP server ecosystem insights useful for production scenarios.
ScaleMCP: Dynamic MCP Tools for LLM Agents (2025)	Lumer et al.	Proposes embedding-guided dynamic tool selection with MCP servers.	Relevant for LLM agents integrating external tools (think candidate databases).
MCP-SafetyBench: Safety Benchmarks for MCP (2025)	Zong et al.	Introduces safety evaluation framework for MCP servers with real workflows.	Useful for evaluating job rec systems that leverage MCP servers.

Table: 1

Generative Job Recommendation using Large Language Models [1] introduces a Generative Job Recommendation framework that leverages large language models to directly generate personalized job recommendations rather than merely ranking predefined listings. The system utilizes transformer-based LLMs to semantically encode candidate resumes and job descriptions, enabling deeper contextual matching between skills, experience, and job requirements. Unlike traditional collaborative or content-based filtering approaches, this model addresses the cold-start problem by relying on natural language understanding rather than historical interaction data. It also integrates explainable AI mechanisms, allowing the model to justify why a particular job suits a candidate. The generative approach enhances personalization and adaptability in dynamic labor markets. This technology represents a shift from retrieval-based to language-driven recommendation systems.

Large Language Models as Zero-Shot Rankers for Recommendation Systems [2] explores the capability of large language models to act as zero-shot rankers in recommendation systems without task-specific training. The technology employs pretrained LLMs to score and rank items, such as job postings, based on natural language prompts describing user preferences. By eliminating the need for labelled training data, the system significantly reduces development and deployment complexity. The ranking decisions rely on the reasoning and world knowledge embedded within LLMs, making them suitable for domain-agnostic recommendation tasks. However, prompt engineering plays a critical role in performance optimization. This approach demonstrates the feasibility of using general-purpose LLMs as ranking engines in job recommendation pipelines.

Chat-REC: Interactive and Explainable LLM-Augmented Recommendation System [3] presents an interactive recommendation architecture that integrates LLMs into

conversational recommender systems. The technology allows users to refine job preferences through natural language dialogue, improving recommendation relevance over multiple interaction rounds. It combines traditional recommender backends with LLM-powered reasoning layers to generate explanations and follow-up questions. The system enhances transparency by clearly articulating why certain jobs are suggested, increasing user trust. This conversational paradigm supports cross-domain reasoning and dynamic preference updates. Overall, Chat-REC demonstrates how LLM-driven conversational interfaces can significantly enhance user engagement in job recommendation platforms.

RecRanker: Instruction-Tuned LLMs for Recommendation Ranking [4] introduces an instruction-tuned large language model designed specifically for ranking tasks in recommender systems. The technology fine-tunes LLMs using natural language instructions to align ranking outputs with recommendation objectives such as relevance and fairness. By framing ranking as an instruction-following task, the model improves controllability and interpretability compared to black-box neural rankers. It supports hybrid ranking scenarios where structured features and unstructured text are jointly considered. This approach reduces dependency on handcrafted ranking functions and allows easier adaptation

across domains. Rec Ranker highlights the effectiveness of instruction-based fine-tuning for scalable job recommendation systems.

The traditional job recommendation systems, when compared with job recommendation approaches based on Large Language Models (LLMs) and Model Context Protocol (MCP) servers, differ mainly in their ability to understand and reason over unstructured data. Conventional systems rely on collaborative filtering, content-based filtering, or hybrid techniques that require structured features and historical user-item interactions. In contrast, LLM-based job recommendation systems leverage deep semantic understanding of resumes, job descriptions, and skill requirements to generate or rank suitable job opportunities. MCP servers act as an integration layer that enables LLMs to dynamically access external tools, databases, and real-time job market information during inference. The recommended jobs are produced through prompt-driven reasoning and contextual retrieval, and their relevance is evaluated using recommendation metrics such as precision, recall, and ranking scores. This literature review provides an overview of the existing LLM-driven job recommendation techniques and MCP-enabled architectures that enhance scalability, adaptability, and explainability in intelligent recruitment systems.

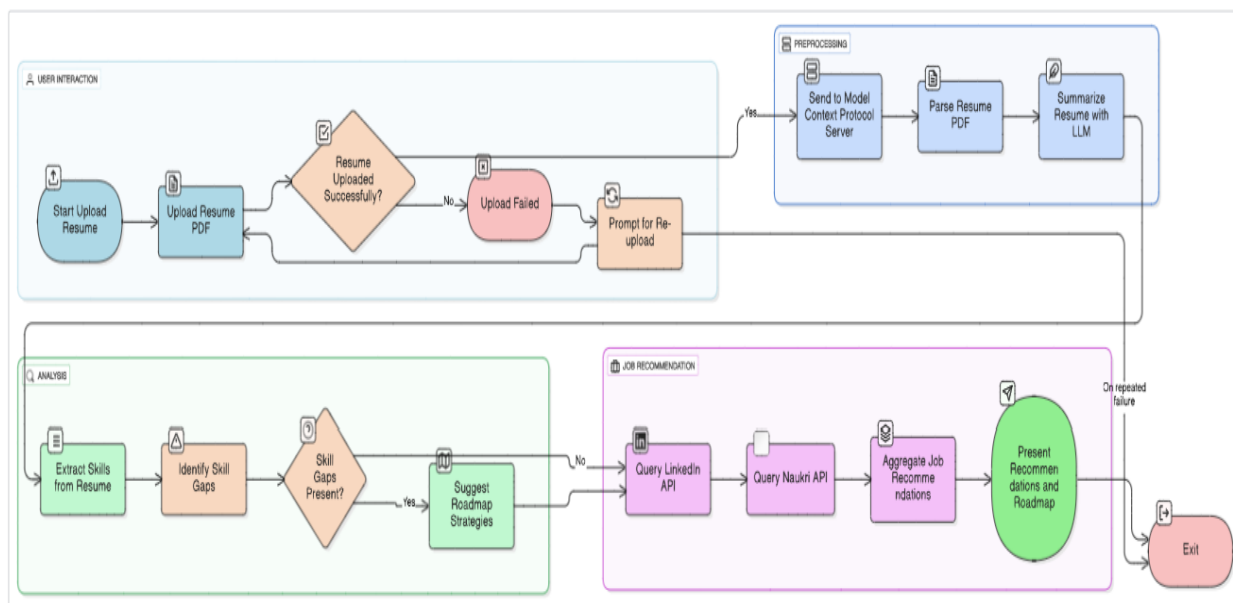


Figure 1: Workflow of Job Recommendation System

Conclusion

This literature survey examined recent advancements in AI-based job recommendation systems that leverage Large Language Models (LLMs) and Model Context Protocol (MCP) servers to overcome the limitations of traditional recommendation techniques. The reviewed studies demonstrate that LLMs significantly improve semantic understanding of resumes and job descriptions, enabling more accurate, personalized, and explainable job recommendations. MCP servers further enhance these systems by facilitating seamless integration of external data sources, tools, and real-time job market information during model inference. The combination of LLM reasoning and MCP-based tool orchestration supports dynamic, scalable, and context-aware recommendation workflows. However, challenges such as model stability, bias, evaluation consistency,

issues, and challenges. Journal of Big Data. Classical RS review useful for baseline comparison with LLM methods.

[17] Wu, L., Qiu, Z., Zheng, Z., Zhu, H., & Chen, E. (2024). Exploring Large Language Models for Graph Data Understanding in Online Job Recommendations. arXiv preprint. Proposes using LLMs to understand and leverage behaviour graphs for more accurate job recommendations.

[18] Hou, X., Zhao, Y., Wang, S., & Wang, H. (2025). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions.

[19] Song, H., Shen, Y., Luo, W., Guo, L., Chen, T., Wang, J., Li, B., Zhang, X. (2025). Beyond the Protocol: Unveiling Attack Vectors in the Model Context Protocol Ecosystem.

[20] Zeng, G., Mao, J., Liu, B., Geng, N., Shang, X., & Gao, Q. (2025). Using the Model Context Protocol (MCP) for Intent-Based Network Troubleshooting Automation.

References

- [1] Generative Job Recommendation using Large Language Models by authors Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, Hui Xiong. 2023
- [2] Large Language Models are Zero-Shot Rankers for Recommender Systems (2023) by author Hou et al.
- [3] Chat-REC: Interactive & Explainable LLM-Augmented Recommender (2023) by authors Gao, Sheng, Xiang et al.
- [4] RecRanker: Instruction-Tuned LLM as Ranker (2023) by authors Luo, He, Zhao et al.
- [5] Large Language Models are Not Stable Recommender Systems (2023) by authors Ma, Cheng, Zhu, Xiong.
- [6] LLM4Jobs: Unsupervised Occupation Extraction (2023) by authors Li, Kang, De Bie.
- [7] Enhance Large Language Models as Recommendation Systems (2025) by authors (Critic-LLM-RS).
- [8] Integrating LLMs with Near-Real-Time Data for Job Rec (2024).
- [9] Balancing Efficiency & Accuracy in LLM Job Matching (2025) by Job Pilot paper.
- [10] Enhancing E-Recruitment via Text Summarization & LLMs (2025) by authors MDPI study.
- [11] Job Recommender Systems: SLR & Challenges (2025) by authors Springer Big Data Review.
- [12] Recommender Systems Meet LLM Agents: Survey (2025) by authors Zhu et al.
- [13] A Measurement Study of Model Context Protocol (2025) by authors Guo et al.
- [14] Scale MCP: Dynamic MCP Tools for LLM Agents (2025) by authors Lumer et al.
- [15] MCP-Safety Bench: Safety Benchmarks for MCP (2025) by authors Zong et al.
- [16] Saito, Y. & Sugiyama, K. (2025). Job recommender systems: a systematic literature review, applications, open

