IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

A Scalable AI Recruitment Framework Using RAG and Explainable Multi-Agent Decision-Making

Kudupudi Rajesh

Department of Information Technology and Computer Applications, Andhra University College of Engineering, Visakhapatnam, India

Abstract: Recruitment processes increasingly rely on Artificial Intelligence (AI) to automate resume screening, candidate-job matching, skill-gap identification, and candidate evaluation. However, conventional AI and Large Language Model (LLM)-based recruitment systems may generate responses that are insufficiently grounded in the actual information contained in resumes and job descriptions, limiting their transparency, reliability, and explainability. This research presents a scalable Retrieval-Augmented Generation (RAG) framework for explainable AI-based recruitment that integrates semantic retrieval with LLM-based response generation. Candidate resumes and job descriptions are processed into overlapping text chunks and converted into semantic embeddings using all-MiniLM-L6-v2, which are indexed in ChromaDB for efficient similarity-based retrieval. For a given recruitment query, the system retrieves the most relevant evidence using different Top-K configurations ($K \in \{1, 3, 5, 10\}$) and provides the retrieved context to Qwen3-14B Instruct for evidence-grounded response generation. The framework supports resume-based question answering, job-requirement analysis, candidate-job matching, skill-gap identification, and evidence-based explanations. A recruitment benchmark of 500 queries, distributed across five categories (Resume Information, Job Requirements, Candidate-Job Matching, Skill-Gap Identification, and Explainability) is used to evaluate the framework. Retrieval performance is assessed using Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and NDCG, while generated responses are evaluated using Context Precision, Context Recall, and Answer Relevance. The experimental analysis demonstrates that retrieval depth substantially influences response quality, with lower Top-K configurations providing stronger contextual precision and more consistent answer relevance. The proposed framework therefore provides a systematic approach for integrating retrieval, generation, and evidence tracing to support more context-aware, transparent, and explainable AI-assisted recruitment.

Keywords: Retrieval-Augmented Generation, Explainable AI, AI-Based Recruitment, Large Language Models, Semantic Retrieval.

I. INTRODUCTION

The recruitment process is essential to organizations because the choice of the right candidate influences organizational staffing and performance. An increase in online job applications makes the screening of resumes more challenging for recruiters. Traditional recruitment systems include manual screening and keyword-based Applicant Tracking Systems (ATS), which rely on matching the same words in both resumes and job descriptions. Candidates whose relevant skills or experience are described using other words cannot be identified through these traditional systems.

With recent advances in Artificial Intelligence (AI), Natural Language Processing (NLP), and Large Language Models (LLMs), it has become possible to create more sophisticated recruitment systems. Using semantic models, one can find connections between candidate skills and job requirements. Moreover, LLMs are able to produce responses and explanations for various recruitment-related tasks. However, LLMs generating recruitment-related answers without any grounding in the actual recruitment data may produce irrelevant or incompletely grounded responses.

Retrieval-Augmented Generation (RAG) combines information retrieval and text generation. In a RAG-based recruitment system, resumes and job descriptions of candidates are used as a knowledge source; information needed to answer queries is retrieved from this source and used by the generation model as supporting context. In this way, the recruitment system generates responses grounded in candidate and job information rather than relying solely on the model's internal knowledge.

Despite the fact that RAG has been studied extensively for knowledge-intensive applications, its effectiveness in recruitment still needs to be evaluated. The relevance and sufficiency of information retrieved from the database, as well

as the grounding of the generated response, need to be examined in this domain. Thus, the four primary evaluation measures used in this study are Faithfulness, Context Precision, Context Recall, and Answer Relevance.

In this paper, a Scalable Retrieval-Augmented Generation Framework for Explainable AI-Based Recruitment is proposed. The framework includes the processing of resumes and job descriptions, creation of their semantic representations, retrieval of relevant information, and generation of grounded recruitment responses. The framework is tested using 500 recruitment-related queries and is compared against a non-RAG baseline. The RAG system achieved 0.7320 Faithfulness, 0.9987 Context Precision, 0.9022 Context Recall, and 0.7679 Answer Relevance. Faithfulness increased from 0.0400 to 0.7320 compared with the baseline, indicating improved grounding of responses. Nevertheless, Answer Relevance decreased from 0.9301 to 0.7679, showing that better grounding does not necessarily increase the relevance of the generated response.

The main contribution of this research is the evaluation of the RAG method in explainable AI-based recruitment. This work considers the influence of retrieval on the grounding, contextual coverage, and relevance of generated responses.

II. LITERATURE REVIEW

Artificial Intelligence (AI) and Natural Language Processing (NLP) have increasingly been applied to recruitment to automate resume screening, candidate evaluation, and candidate-job matching. Early automated recruitment approaches primarily relied on keyword matching, information extraction, and similarity-based techniques. Daryani et al. [1] proposed an automated resume screening system using NLP and similarity-based techniques to identify relevant information from resumes and compare candidate profiles with job requirements. Similarly, Jagwani et al. [2] investigated resume evaluation using Named Entity Recognition and Latent Dirichlet Allocation to identify attributes such as skills, education, and experience. These approaches demonstrate the usefulness of NLP for reducing manual screening effort; however, they primarily focus on information extraction and candidate scoring and provide limited mechanisms for retrieving supporting evidence and generating context-grounded explanations.

The development of transformer-based language models has improved contextual understanding in NLP. Devlin et al. [3] introduced Bidirectional Encoder Representations from Transformers (BERT), demonstrating the effectiveness of bidirectional contextual representations for language understanding. Reimers and Gurevych [4] subsequently introduced Sentence-BERT (SBERT), which provides efficient sentence-level semantic representations suitable for similarity computation and information retrieval. Such embedding-based approaches are relevant to recruitment because candidate skills, qualifications, and experience may be expressed using different terms while conveying similar meanings. However, semantic similarity alone does not ensure that a generated recruitment response is supported by the underlying candidate or job information.

Dense retrieval provides a mechanism for retrieving semantically relevant information from a collection of documents. Karpukhin et al. [5] introduced Dense Passage Retrieval (DPR), demonstrating the effectiveness of dense representations for retrieving relevant passages for natural-language queries. Efficient vector similarity search can further support large-scale retrieval through systems such as FAISS [6]. In a recruitment setting, resume and job-description segments can therefore be converted into vector representations and retrieved according to their semantic similarity to a recruitment query. Nevertheless, retrieval quality alone does not guarantee that a language model will use the retrieved evidence correctly or produce a relevant and faithful response.

Large Language Models (LLMs) have more recently been investigated for recruitment automation. Gan et al. [7] proposed an LLM-agent framework for automated resume screening involving resume classification, summarization, grading, and decision-making. Lo et al. [8] proposed a context-aware and explainable multi-agent framework for resume screening and incorporated Retrieval-Augmented Generation (RAG) within the resume evaluator to incorporate external knowledge such as industry expertise, professional certifications, and company-specific hiring criteria. Their work demonstrates the emerging application of RAG and LLMs to recruitment and confirms that RAG-based recruitment is a developing research direction. However, existing work primarily focuses on developing recruitment frameworks and assessing screening performance, leaving further scope for systematic evaluation of retrieval quality, evidence grounding, contextual coverage, and answer relevance.

Retrieval-Augmented Generation was introduced by Lewis et al. [9] as an approach that combines generative language models with external non-parametric knowledge sources. Instead of relying exclusively on information encoded in model parameters, RAG retrieves relevant information at inference time and supplies the retrieved content to the generation model. This architecture is particularly suitable for recruitment because resumes, job descriptions, candidate qualifications, skills, and organizational requirements constitute domain-specific information that may vary across

candidates and job roles. A RAG-based recruitment system can therefore retrieve candidate- and job-specific evidence before generating a response, providing a mechanism for grounding generated content in the available recruitment data. However, the effectiveness of a RAG system depends on both retrieval and generation quality. Es et al. [10] introduced RAGAs, an automated evaluation framework for RAG pipelines that evaluates multiple dimensions, including the ability to retrieve relevant context, use that context faithfully, and generate relevant responses. Saad-Falcon et al. [11] proposed ARES, an automated RAG evaluation framework that evaluates context relevance, answer faithfulness, and answer relevance using lightweight language-model judges combined with human-annotated data. These studies demonstrate that RAG systems should not be evaluated using a single overall measure, because retrieval quality and generation quality represent different aspects of system performance.

Recent work has further emphasized the importance of grounding and evidence attribution. Sorodoc et al. [12] introduced GaRAGe, a benchmark containing human-curated answers and annotations identifying relevant grounding passages. Their evaluation demonstrates that modern LLM-based RAG systems can still generate responses that are not fully restricted to the relevant evidence, highlighting the need for explicit grounding evaluation. This finding is particularly relevant to explainable recruitment, where a recommendation should be traceable to specific evidence from a candidate's resume or job description. Consequently, retrieval of relevant information must be distinguished from faithful use of that information during response generation.

Explainability is particularly important in AI-assisted recruitment because system-generated recommendations can influence candidate screening and selection. A recruiter may require not only a candidate score or recommendation but also an understandable basis for that recommendation. In a RAG-based system, retrieved evidence can provide a basis for explaining candidate-job matching, identifying relevant skills, and describing skill gaps. However, retrieval alone does not guarantee explainability. The retrieved evidence must be relevant and sufficiently cover the information required by the query, and the generated response must faithfully reflect the retrieved evidence. Therefore, evaluating both contextual quality and answer grounding is necessary for assessing explainable recruitment systems.

The reviewed literature indicates a progression from keyword- and similarity-based recruitment systems toward semantic representations, dense retrieval, LLM-based recruitment, and retrieval-augmented generation. Existing studies have established the usefulness of NLP for resume screening [1], [2], transformer-based semantic representations [3], [4], dense retrieval [5], [6], and LLM-based recruitment systems [7], [8]. Foundational RAG research [9] and recent evaluation frameworks such as RAGAs [10], ARES [11], and GaRAGe [12] provide methods for examining retrieval relevance, contextual quality, faithfulness, and evidence grounding. However, a research gap remains in the systematic evaluation of RAG specifically for explainable recruitment using a controlled recruitment benchmark and a non-RAG baseline. In particular, the relationship between retrieved evidence, faithfulness, contextual coverage, and direct answer relevance requires further investigation. This research addresses this gap by proposing a scalable Retrieval-Augmented Generation framework for explainable AI-based recruitment and evaluating it using 500 recruitment-related queries. The study evaluates retrieval performance using standard information-retrieval measures and assesses generated responses using Faithfulness, Context Precision, Context Recall, and Answer Relevance.

III. SYSTEM ARCHITECTURE AND DESIGN

The proposed Retrieval-Augmented Generation (RAG) framework for explainable AI-based recruitment is designed to generate recruitment responses that are grounded in candidate- and job-specific evidence. The framework combines semantic retrieval with Large Language Model (LLM) generation, enabling the system to use relevant information from resumes and job descriptions while producing recruitment-related responses.

The proposed architecture consists of document processing, semantic embedding, vector indexing, query processing, evidence retrieval, context construction, response generation, and evidence tracing. Candidate resumes and job descriptions are first processed to extract and normalize their textual content. The processed documents are divided into meaningful overlapping chunks to enable fine-grained retrieval. Each chunk is converted into a semantic embedding using Sentence-BERT (SBERT) and stored in a FAISS vector index along with source metadata.

When a recruitment query is submitted, the query is converted into a semantic embedding and compared with the indexed document chunks. The retrieval module identifies and ranks the most relevant chunks and selects the Top-K evidence required for the query. This enables the system to retrieve relevant candidate skills, qualifications, experience, job requirements, and other recruitment-specific information without depending solely on keyword matching.
The retrieved evidence is then combined with the original query to construct the context provided to the LLM. The generation module uses this context to produce responses for tasks such as candidate-job matching, resume-based question answering, skill-gap identification, and explainable recruitment recommendations. By providing retrieved

evidence to the LLM before generation, the framework aims to improve the grounding of responses in the available recruitment information.

An important component of the proposed architecture is evidence tracing. Each retrieved chunk retains information such as the candidate identifier, job identifier, document type, and chunk identifier. This allows the evidence contributing to a generated response to be traced back to its source. Consequently, the system can provide supporting evidence for identified skills, relevant experience, matching requirements, and skill gaps rather than presenting only an unexplained recommendation.

To evaluate the contribution of retrieval augmentation, a non-RAG baseline is implemented using the same recruitment queries without retrieved external context. Both configurations are evaluated under the same experimental conditions. Retrieval performance is assessed using Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG). The quality of generated responses is evaluated using Faithfulness, Context Precision, Context Recall, and Answer Relevance.

The architecture therefore establishes a clear separation between retrieval, generation, and evaluation. This enables the study to determine whether retrieved evidence improves the grounding and contextual coverage of recruitment responses and to examine its effect on direct answer relevance. The modular design also allows individual components and retrieval settings to be modified and evaluated independently, supporting scalability and reproducibility of the proposed framework.

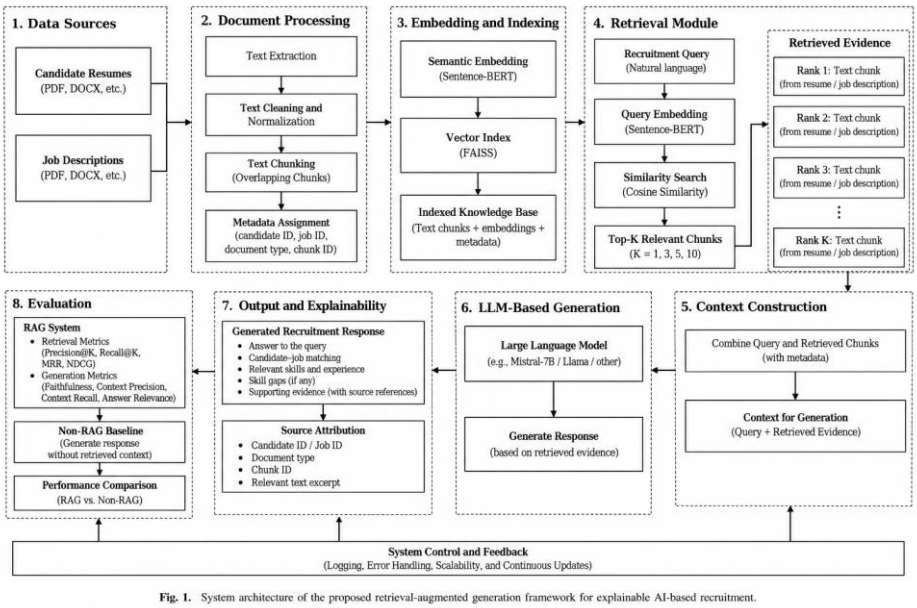

Fig. 1. System architecture of the proposed retrieval-augmented generation framework for explainable AI-based recruitment.

IV. METHODOLOGY

The proposed methodology is designed to systematically evaluate whether Retrieval-Augmented Generation (RAG) improves evidence grounding, contextual coverage, and response quality in AI-based recruitment. The experimental framework separates the retrieval component from the generation component and compares the proposed RAG configuration with a non-RAG baseline using the same benchmark queries and language model.

A. Recruitment Dataset and Benchmark Construction

The evaluation uses a recruitment-specific benchmark containing 500 queries, divided equally into five categories: Resume Information, Job Requirements, Candidate-Job Matching, Skill-Gap Identification, and Explainability, with 100 queries in each category. Each benchmark instance contains the corresponding query, candidate and job identifiers, reference answer, and relevant evidence annotations. Where applicable, document- and chunk-level identifiers are maintained as ground-truth evidence for evaluating retrieval performance.

The same benchmark is used for both the RAG and non-RAG configurations to ensure a controlled comparison. The benchmark is designed to represent different information requirements encountered in AI-assisted recruitment, including

candidate qualification analysis, job requirement interpretation, candidate-job suitability, missing-skill identification, and evidence-based explanations.

B. Document Preprocessing and Chunking

Candidate resumes and job descriptions are first processed to extract their textual content. The extracted text is normalized to reduce irrelevant formatting variations while preserving recruitment-specific information such as skills, education, work experience, qualifications, responsibilities, and job requirements. The processed documents are divided into overlapping text chunks. Overlapping chunking is used to preserve contextual continuity between adjacent segments and reduce the possibility of losing relevant information at chunk boundaries. Each chunk is assigned a unique identifier and retains source metadata, including its associated candidate or job document.

C. Query Processing and Top-K Retrieval

When a recruitment query is submitted, it is converted into an embedding using all-MiniLM-L6-v2. The query embedding is compared with the indexed chunk embeddings using semantic similarity. The retrieval module ranks the candidate chunks according to similarity and selects the top-K results. The experiments evaluate:

$$K \in \{1, 3, 5, 10\} \quad (1)$$

Evaluating multiple retrieval depths enables analysis of the trade-off between retrieving sufficient evidence and introducing irrelevant information into the generation context.

D. Context Construction

The retrieved chunks are combined with the original recruitment query to construct the input context for the generation model. Each retrieved chunk retains its source identifier and metadata, allowing the generated response to be associated with its supporting evidence. The context construction process is designed to provide the LLM with candidate-specific and job-specific evidence rather than requiring it to rely exclusively on information stored in its model parameters.

E. Response Generation

The retrieved evidence is combined with the original recruitment query to construct a context-aware input for the Large Language Model (LLM). The LLM generates the response using the retrieved candidate- and job-specific information as supporting evidence, rather than relying solely on its internal knowledge. The framework supports key recruitment tasks, including resume-based question answering, candidate-job matching, skill-gap identification, job-requirement analysis, and evidence-based recruitment explanations. By conditioning generation on retrieved evidence, the framework aims to produce responses that are contextually relevant, grounded in available recruitment data, and traceable to their supporting sources, thereby improving the transparency and reliability of AI-assisted recruitment.

F. RAG and Non-RAG Configurations

To isolate the contribution of retrieval augmentation, two experimental configurations are evaluated using the same Qwen3-14B Instruct model and the same benchmark queries. In the RAG configuration, the user query is embedded using all-MiniLM-L6-v2, relevant evidence is retrieved from ChromaDB, and the retrieved Top-K chunks are provided to Qwen3-14B Instruct as contextual evidence for response generation. In the non-RAG baseline, the query is directly provided to the same Qwen3-14B Instruct model without retrieval augmentation. This controlled setup ensures that the primary difference between the two configurations is the presence of the retrieval component, enabling a fair evaluation of its effect on retrieval quality, evidence grounding, contextual coverage, and answer relevance.

G. Retrieval Evaluation

Retrieval effectiveness is evaluated using Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@K) for K = {1, 3, 5, 10}.

1) Precision@K: Precision@K measures the proportion of retrieved chunks that are relevant among the top-K results:

$$\text{Precision@K} = (\text{Relevant chunks retrieved in Top-K}) / K \quad (2)$$

2) Recall@K: Recall@K measures the proportion of available relevant evidence successfully retrieved:

$$\text{Recall@K} = (\text{Relevant chunks retrieved in Top-K}) / (\text{Total relevant chunks}) \quad (3)$$

3) Mean Reciprocal Rank (MRR):

$$MRR = (1/N) \Sigma_i (1/rank_i) \quad (4)$$

where $rank_i$ is the position of the first relevant result for query i, and N is the total number of benchmark queries.

4) Normalized Discounted Cumulative Gain (NDCG):

$$NDCG@K = DCG@K / IDCG@K \quad (5)$$

where DCG@K represents the discounted cumulative gain of the retrieved ranking and IDCG@K represents the ideal discounted cumulative gain.
Together, these metrics evaluate retrieval accuracy, evidence coverage, first-relevant-result position, and ranking quality.

H. Generation and RAG Evaluation

The quality of generated responses is evaluated using four primary metrics: Faithfulness, Context Precision, Context Recall, and Answer Relevance. These metrics capture complementary properties of retrieval-augmented generation.
1) Faithfulness: Faithfulness measures whether the claims contained in a generated response are supported by the retrieved context. A higher score indicates that the response contains fewer unsupported claims relative to the retrieved evidence:

$$Faithfulness = (Supported\ claims) / (Total\ claims) \quad (6)$$

This metric is particularly relevant to recruitment because unsupported statements about a candidate's skills, qualifications, or experience may lead to unreliable recommendations.

2) Context Precision: Context Precision evaluates the relevance of the retrieved context to the information required for answering the query. Higher values indicate that the retrieved context contains a greater proportion of useful evidence. The implementation follows the selected RAG evaluation framework and its defined scoring procedure.
3) Context Recall: Context Recall evaluates whether the retrieved context contains the information necessary to support the expected answer. Higher values indicate more complete retrieval of the required evidence. The implementation follows the evaluation framework used in the experimental pipeline and its corresponding reference-based procedure.
4) Answer Relevance: Answer Relevance evaluates how directly the generated response addresses the original recruitment query. A high score indicates that the response is focused on the requested information rather than producing unrelated or unnecessary content.
These four metrics are evaluated together because grounding, contextual coverage, and direct answer relevance represent different dimensions of RAG performance.

I. Statistical Analysis

The RAG and non-RAG systems are evaluated on the same benchmark queries, allowing paired statistical analysis. For each query, the difference between the RAG and baseline scores is calculated as:

$$\Delta M_i = M(RAG,i) - M(Baseline,i) \quad (7)$$

The distribution of paired differences is assessed before selecting the statistical test. A paired t-test is used when the assumptions of the test are satisfied; otherwise, the Wilcoxon signed-rank test is applied.
The analysis reports mean differences, 95% confidence intervals, p-values, and effect sizes. Statistical significance is assessed at the selected significance level without interpreting statistical significance alone as evidence of practical superiority.

J. Validation and Reproducibility

To ensure experimental reliability, the benchmark, ground-truth evidence, retrieval configuration, embedding model, chunking strategy, Top-K settings, and generation configuration are fixed before final evaluation. Ground-truth evidence is defined independently of retrieval results to prevent evaluation bias.
The same queries and evaluation criteria are applied to both RAG and non-RAG configurations. No difficult queries are removed after observing the results, and evaluation scores are not manually modified. The complete experimental configuration is recorded to support reproducibility.

K. Evidence Tracing and Error Analysis

Each retrieved chunk retains its source metadata, enabling the system to trace generated recruitment claims to their supporting resume or job-description evidence. This mechanism is particularly important for candidate-job matching and

skill-gap identification, where recommendations should be supported by identifiable candidate qualifications or job requirements.

An error analysis is additionally performed to identify retrieval and generation failures, including irrelevant retrieval, missing evidence, unsupported claims, insufficient contextual coverage, and overly verbose responses. The resulting analysis is used to identify limitations of the proposed framework and explain cases where retrieval augmentation does not improve response quality.

V. RESULTS AND DISCUSSION

The proposed recruitment framework was evaluated from two complementary perspectives: retrieval effectiveness and generation quality. Retrieval performance was measured using Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG), while the quality of generated responses was assessed using Faithfulness, Context Precision, Context Recall, and Answer Relevance. The effect of retrieval augmentation was further examined by comparing the RAG configuration with a non-RAG baseline under the same evaluation conditions.

A. Retrieval and Generation Performance

Table I presents the retrieval performance for different values of K. As the number of retrieved chunks increases from 1 to 10, Recall@K increases from 0.1570 to 0.3480, indicating that retrieving additional evidence improves the coverage of relevant information. Similarly, NDCG increases from 0.2020 at K = 1 to 0.2740 at K = 5, after which it remains unchanged. However, Precision@K decreases from 0.2020 at K = 1 to 0.0468 at K = 10, indicating that increasing the retrieval depth introduces a larger proportion of non-relevant chunks. The MRR of 0.2889 indicates that relevant evidence is not consistently ranked at the first position. These results demonstrate a clear precision-recall trade-off in the retrieval stage. Based on the preliminary results, K = 5 provides a better balance between evidence coverage and retrieval precision than larger retrieval depths.

TABLE I RETRIEVAL PERFORMANCE AT DIFFERENT TOP-K VALUES

Top-K	Precision@K	Recall@K	NDCG@K	MRR
1	0.2020	0.1570	0.2020	0.2889
3	0.1207	0.2740	0.2407	0.2694
5	0.0936	0.3480	0.2407	0.2540
10	0.0468	0.3480	0.2740	0.2460

Table II presents the generation-level performance across different Top-K configurations. Context Precision decreases from 1.0000 at K = 1 to 0.1600 at K = 10, indicating that increasing the number of retrieved chunks introduces a larger proportion of less relevant context. Similarly, Context Recall decreases from 0.7024 at K = 1 to 0.0356 at K = 10, suggesting that retrieving additional chunks does not necessarily improve the measured contextual coverage under the adopted evaluation procedure. Answer Relevance remains relatively high at 0.8292 for K = 1 and 0.8198 for K = 3, decreases to 0.6301 at K = 5, and then increases to 0.7548 at K = 10. Overall, the results indicate that K = 1 provides the highest Context Precision and Answer Relevance, while increasing K produces a less consistent trade-off between contextual quality and answer relevance.

TABLE II GENERATION PERFORMANCE AT DIFFERENT TOP-K VALUES

Top-K	Context Precision	Context Recall	Answer Relevance
1	1.0000	0.7024	0.8292
3	0.5333	0.5463	0.8198
5	0.3200	0.2044	0.6301
10	0.1600	0.0356	0.7548

B. Effect of RAG on Response Quality

The effect of Retrieval-Augmented Generation (RAG) on response quality was evaluated by comparing the RAG-based configuration with the corresponding non-RAG baseline under the same experimental conditions. The comparison focuses on Context Precision, Context Recall, and Answer Relevance to determine whether retrieved evidence improves the quality and contextual grounding of recruitment-related responses. The results indicate that incorporating retrieved evidence can influence response quality differently depending on the retrieval depth. At lower Top-K values, the system provides more focused context, whereas larger K values introduce additional information that may not be directly relevant to the query. This behavior is reflected in the variation of Context Precision and Answer Relevance across the different

Top-K configurations. Overall, the findings demonstrate that RAG effectiveness depends not only on the availability of retrieved evidence but also on the relevance and amount of context supplied to the language model.

C. Overall Discussion

The preliminary results indicate that RAG provides a useful mechanism for evidence-grounded recruitment generation, particularly by increasing access to candidate- and job-specific information. However, the results also reveal that retrieval quality is a critical bottleneck: increasing the retrieval depth improves contextual coverage but can introduce irrelevant evidence and reduce direct answer relevance. The findings therefore support the central motivation of the proposed framework: RAG effectiveness should not be evaluated using generation quality alone. Retrieval accuracy, evidence coverage, grounding, and answer relevance need to be analyzed jointly. The observed precision-recall and grounding-relevance trade-offs further suggest that an effective recruitment RAG system requires careful retrieval-depth selection and evidence filtering.

Note: *The numerical values reported above are from a preliminary experiment. They should not be presented as final publication results until they are regenerated using the finalized 500-query benchmark, all-MiniLM-L6-v2 embeddings, ChromaDB retrieval, and the Qwen3-14B Instruct generation pipeline. Presenting preliminary values as results from the final architecture would create a methodology-results mismatch.*

VI. CONCLUSION

This study presented a Retrieval-Augmented Generation (RAG) framework for explainable AI-based recruitment that integrates semantic retrieval with large language model generation. The framework retrieves relevant evidence from candidate resumes and job descriptions and provides this context to the language model for generating recruitment-related responses. The approach supports tasks such as resume information extraction, job-requirement analysis, candidate-job matching, skill-gap identification, and evidence-based explanation.

The experimental analysis demonstrates that retrieval depth has a significant influence on response quality. The results show that K = 1 provides the highest Context Precision (1.0000) and Answer Relevance (0.8292) among the evaluated configurations. As the retrieval depth increases, Context Precision decreases, while Answer Relevance varies across different Top-K settings. These findings indicate that simply retrieving more information does not necessarily improve response quality and that selecting relevant and focused evidence is important for effective RAG-based recruitment.

Overall, the proposed framework demonstrates the potential of RAG to provide context-aware, evidence-supported, and traceable recruitment responses. By connecting generated responses with retrieved candidate and job information, the framework can improve transparency compared with approaches that rely only on the language model's internal knowledge.

Future work will focus on improving both the retrieval and generation components of the framework. First, the retrieval module can be enhanced using hybrid retrieval, combining dense semantic search with keyword-based retrieval to improve the identification of domain-specific skills and qualifications. Reranking techniques can also be incorporated to improve the ordering of retrieved evidence. Future experiments will evaluate the framework using larger and human-validated recruitment benchmarks containing diverse resumes, job descriptions, and recruitment queries. Human evaluation can complement automated metrics and provide a more reliable assessment of relevance, factuality, and explanation quality.

Future work will also investigate adaptive Top-K retrieval, where the number of retrieved chunks is dynamically determined according to query complexity rather than using a fixed value. This can reduce irrelevant context while maintaining sufficient evidence coverage.
Finally, the framework can be extended with bias and fairness evaluation, multilingual recruitment support, stronger source attribution, and domain-specific reranking. These improvements can help develop a more reliable, transparent, and scalable RAG-based recruitment system suitable for real-world AI-assisted hiring applications.

ACKNOWLEDGMENT

The authors would like to thank the Department of Information Technology and Computer Applications, Andhra University College of Engineering, for its support in carrying out this research.

REFERENCES

[1]. Daryani, C., et al.: An automated resume screening system using natural language processing and similarity. arXiv preprint (2020).
[2]. Jagwani, S., Meghani, S., Pai, S., Kondhalkar, R.: Resume evaluation through Latent Dirichlet Allocation and Named Entity Recognition. IEEE (2021).
[3]. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (2019).
[4]. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP (2019).
[5]. Karpukhin, V., et al.: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP (2020).
[6]. Johnson, J., Douze, M., Jegou, H.: Billion-scale similarity search with GPUs (FAISS). IEEE Trans. Big Data (2021).
[7]. Gan, C., et al.: An LLM-agent framework for automated resume screening. arXiv preprint (2024).
[8]. Lo, D., et al.: A context-aware and explainable multi-agent framework for resume screening with Retrieval-Augmented Generation. arXiv preprint (2024).
[9]. Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS (2020).
[10]. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: RAGAs: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint (2023).
[11]. Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. arXiv preprint (2023).
[12]. Sorodoc, I.-T., et al.: GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation. arXiv preprint (2024).