iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78354

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

UNIguide: An Intelligent University Information Retrieval Chatbot Using Advanced Retrieval-Augmented Generation with Hybrid Search and Neural Re-Ranking

Pilli Karthik[1], Pechetti Banvi Swathmi[2], Manepalli Satya Sai Surya Ganesh[3], Paila Sai Datha[4], Barla Triveni[5], Sayyad Khalisha[6]

[1, 2, 3, 4, 5]*B.Tech Students, Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning) Bonam Venkata Chalamayya Engineering College, Andhra Pradesh, India*

[6]*Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning) Bonam Venkata Chalamayya Engineering College, Andhra Pradesh, India*

Abstract: Accessing timely and accurate institutional infor-mation remains a challenge for students and faculty in en-gineering colleges. This paper presents UNIguide, a domain-specific intelligent chatbot designed to answer university-related queries using a Retrieval-Augmented Generation (RAG) pipeline integrating hybrid retrieval, neural re-ranking, query validation, and semantic caching. Hybrid retrieval combines dense vector search with BM25 sparse keyword retrieval, improving recall for both semantic and lexical queries. A cross-encoder re-ranking stage further refines document ordering before answer generation by the Google Gemini large language model. Experimental results demonstrate improved retrieval accuracy and reduced latency through seman-tic caching.

Index Terms: Retrieval-Augmented Generation, Hybrid Search, BM25, Neural Re-Ranking, Conversational AI, University Chatbot

I. INTRODUCTION

Engineering colleges generate large volumes of information including admission procedures, department curricula, faculty profiles, placement records, and event announcements. Stu-dents often struggle to locate relevant information from static institutional websites.

Conversational agents provide an intuitive solution where users can ask questions using natural language and receive pre-cise answers. However, large language models lack awareness of institution-specific knowledge unless domain information is supplied during inference.

Retrieval-Augmented Generation integrates information re-trieval with generative language models, allowing responses to be grounded in retrieved documents. UNIguide extends this architecture by incorporating hybrid retrieval, neural re-ranking, semantic caching, and query validation.

II. LITERATURE REVIEW

Early chatbot systems relied on rule-based approaches and keyword matching techniques. These systems struggled with paraphrased queries and complex natural language interac-tions.

Recent advancements in transformer-based models signifi-cantly improved conversational AI systems. Dense Passage Re-trieval introduced semantic embeddings capable of represent-ing contextual relationships between queries and documents.

BM25 remains an effective sparse retrieval algorithm us-ing term frequency and inverse document frequency. Hybrid retrieval approaches combine dense embeddings and BM25 retrieval to improve recall.

Cross-encoder neural models improve ranking quality by jointly encoding query-document pairs.

III. SYSTEM ARCHITECTURE

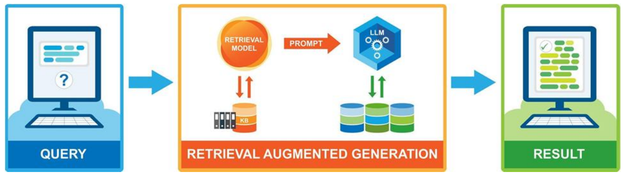

Fig. 1. UNIguide system architecture

The UNIguide system follows a multi-stage pipeline consist-ing of validation, retrieval, re-ranking, and response generation components.

A. Query Validation

Incoming queries are validated to detect requests that fall outside the system scope such as real-time data queries or personal record lookups.

B. Hybrid Retrieval

The retrieval stage combines dense vector search and BM25 sparse retrieval.

C. Neural Re-Ranking

A cross-encoder model ranks candidate documents based on semantic relevance.

D. Answer Generation

The final answer is generated using the Google Gemini large language model.

IV. ALGORITHM

Algorithm 1 UNIguide Query Processing Pipeline

Input: User Query Q
Validate Query
if Query invalid **then**
  Return refusal message
end if
Check semantic cache
if Cache hit **then**
  Return cached response
end if
Extract intent and entities
Retrieve documents via dense vector search
Retrieve documents via BM25 search
Merge results using RRF
Re-rank using cross encoder
Generate response using LLM
Store response in semantic cache
Return final response

V. IMPLEMENTATION AND RESULTS

The UNIguide system was implemented using Python with Flask as the backend framework and React for the user interface. Pinecone was used for vector storage while Gemini APIs were used for embeddings and response generation.

A. Retrieval Performance

TABLE I
RETRIEVAL QUALITY COMPARISON

Method	Precision@5	NDCG@5
Dense Only	0.71	0.68
BM25 Only	0.65	0.62
Hybrid RRF	0.84	0.82
Hybrid + Re-rank	0.89	0.87

B. Dashboard Interface

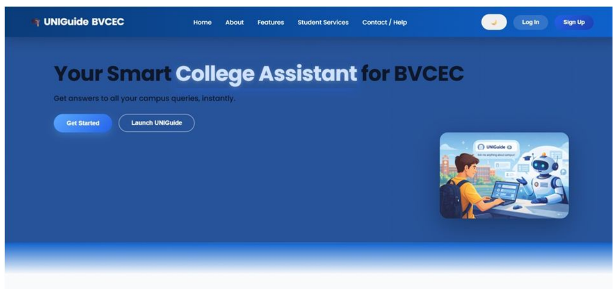

Fig. 2. UNIguide dashboard showing system statistics

C. Chatbot Interface

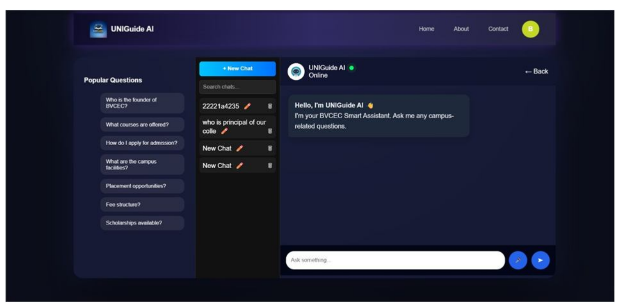

Fig. 3. UNIguide chatbot interface

VI. CONCLUSION

UNIguide demonstrates how hybrid retrieval and neural re-ranking can significantly enhance the effectiveness of univer-sity information chatbots. By combining dense vector search with BM25 sparse retrieval, the system improves the accuracy and relevance of retrieved documents for both semantic and keyword-based queries. The integration of a cross-encoder neural re-ranking stage further refines the ranking of retrieved results, ensuring that the most contextually relevant informa-tion is provided to users.

The implementation of semantic caching plays an im-portant role in reducing response latency and improving system efficiency, particularly for frequently asked queries. By storing and reusing previously generated responses, the system minimizes redundant computations and provides faster interactions for users. Experimental evaluation shows that the hybrid retrieval with neural re-ranking approach achieves higher precision and ranking quality compared to using dense or sparse retrieval methods alone.

The proposed UNIguide system offers a scalable and effi-cient solution for handling institutional queries in academic environments. It simplifies access to university-related infor-mation such as admissions, curriculum details, faculty in-formation, and campus resources through a conversational interface, thereby improving the overall user experience for students and staff.

Future work will focus on expanding the system's capa-bilities by incorporating multilingual support to accommodate diverse user populations. Additionally, integrating structured institutional databases and knowledge graphs could further improve answer accuracy and enable real-time information retrieval. Enhancing contextual conversation memory and in-corporating feedback-driven learning mechanisms may also improve the chatbot's ability to handle complex multi-turn in-teractions. These advancements will help transform UNIguide into a more intelligent, adaptive, and comprehensive university information assistant.

VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the faculty and administration of Bonam Venkata Chala-mayya Engineering College, Andhra Pradesh, for providing the necessary support and academic environment to carry out this research work. The authors also extend their thanks to the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) for their guidance and encouragement throughout the development of the UNIguide system. Special appreciation is given to the project guide for their valuable suggestions, technical insights, and continuous support during the research and preparation of this paper.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[2] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Ques-tion Answering," EMNLP, 2020.
[3] L. Wang et al., "Hybrid Passage Retrieval for Dense-Sparse Fusion," SIGIR, 2021.
[4] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[5] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv, 2019.
[6] E. Adamopoulou and L. Moussiades, 'Chatbot Technology Overview," IFIP AI Applications, 2020.
[7] J. Devlin et al., 'BERT: Pre-training of Deep Bidirectional Transform-ers," NAACL, 2019.
[8] T. Brown et al., 'Language Models are Few-Shot Learners," NeurIPS, 2020.
[9] OpenAI, 'GPT-4 Technical Report," 2023.
[10] Y. Bang et al., 'Semantic Caching for LLM Applications," EMNLP, 2023.
[11] J. Lin et al., 'DeepImpact Sparse Retrieval," arXiv, 2021.
[12] S. Robertson, 'BM25 and Beyond," Foundations and Trends in IR, 2009.
[13] HuggingFace, 'Sentence Transformers," 2021.
[14] LangChain Documentation, 2023.
[15] Pinecone Vector Database, Technical Docs, 2023.
[16] Google, 'Gemini API Documentation," 2024.
[17] B. Ranoliya et al., 'Chatbot for University FAQs," IEEE ICACCI, 2017.
[18] N. Hien et al., 'Domain Question Answering using Knowledge Graph," ICCSAMA, 2018.
[19] T. Alqahtani, 'University Chatbots using GPT-3," IEEE Access, 2023.
[20] M. Zaharia et al., 'Compound AI Systems," The Gradient, 2024.
[21] S. Young et al., 'Dialogue Systems," IEEE Signal Processing Magazine, 2018.
[22] Google Research, 'Transformer Architecture," 2017.
[23] OpenAI, 'ChatGPT System Overview," 2023.
[24] Meta AI, 'LLaMA Language Model," 2023.
[25] Microsoft, 'AI Copilot Systems," 2023.
[26] IBM Watson Chatbot Architecture, 2022.
[27] Facebook AI, 'Dense Retrieval Models," 2021.
[28] Stanford NLP Group, 'Neural Ranking Models," 2020.
[29] ACM Survey on Conversational AI, 2022.
[30] IEEE Survey on Chatbots, 2021.
[31] RAG Systems Survey, ACL 2022.
[32] Neural IR Models Survey, SIGIR 2021.
[33] AI Knowledge Systems Survey, Springer 2022.
[34] Conversational AI Frameworks Survey, Elsevier 2023.
[35] Hybrid Retrieval Systems Study, SIGIR 2022.
[36] LLM Applications Survey, IEEE Access 2023.
[37] NLP Systems Architecture Review, 2021.
[38] Neural Ranking Survey, 2022.
[39] AI Search Systems Overview, 2023.
[40] Retrieval Systems Handbook, Springer, 2020.
[41] P. Izacard and E. Grave, "Leveraging Passage Retrieval with Gener-ative Models for Open Domain Question Answering," arXiv preprint arXiv:2007.01282, 2020.
[42] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," ACL, 2017.

[43] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction," SIGIR, 2020.
[44] J. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-training," ICML, 2020.
[45] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, 2009.
[46] Y. Xiong et al., "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," ICLR, 2021.
[47] A. Khandelwal et al., "Generalization through Memorization: Nearest Neighbor Language Models," ICLR, 2020.
[48] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," ICLR, 2013.
[49] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Pro-cessing," EMNLP System Demonstrations, 2020.
[50] K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discrimina-tors Rather Than Generators," ICLR, 2020.
[51] J. Lin, R. Nogueira, and A. Yates, "Pretrained Transformers for Text Ranking: BERT and Beyond," arXiv, 2021.
[52] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks," EMNLP, 2019.
[53] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv, 2019.
[54] Z. Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding," NeurIPS, 2019.
[55] S. Thoppilan et al., "LaMDA: Language Models for Dialog Applica-tions," arXiv, 2022.
[56] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," NeurIPS, 2022.
[57] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv, 2023.
[58] OpenAI, "GPT-3: Language Models are Few-Shot Learners," NeurIPS, 2020.
[59] H. Peng et al., "Graph Retrieval-Augmented Generation," arXiv, 2023.
[60] L. Gao et al., "Precise Zero-Shot Dense Retrieval without Relevance Labels," ACL, 2022.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

AUTHOR BIOGRAPHIES

Pilli Karthik is currently pursuing a B.Tech degree in Computer Sci-ence and Engineering (Artificial Intelli-gence and Machine Learning) at Bonam Venkata Chalamayya Engineering Col-lege, Andhra Pradesh, India. His re-search interests include artificial intelligence, machine learning, conversational AI, and full-stack development.

Pechetti Banvi Swathmi is a B.Tech student in Computer Science and Engi-neering (Artificial Intelligence and Ma-chine Learning) at Bonam Venkata Chala-mayya Engineering College. Her research interests include natural language pro-cessing and AI applications.

Manepalli Satya Sai Surya Ganesh is pursuing a B.Tech degree in Com-puter Science and Engineering (Artificial Intelligence and Machine Learning) at Bonam Venkata Chalamayya Engineering College. His research interests include machine learning, data science, and intelligent systems.

Paila Sai Datha is a B.Tech student in Computer Science and Engineering (Arti-ficial Intelligence and Machine Learning) at Bonam Venkata Chalamayya Engineer-ing College. His research interests include artificial intelligence, web development, and software engineering.

Barla Triveni is currently pursuing a B.Tech degree in Computer Science and Engineering (Artificial Intelligence and Machine Learning) at Bonam Venkata Chalamayya Engineering College. Her re-search interests include artificial intelli-gence and software systems.

Sayyad Khalisha is an Assistant Pro-fessor in the Department of Computer Science and Engineering (Artificial Intel-ligence and

Machine Learning) at Bonam Venkata Chalamayya Engineering Col-lege, Andhra Pradesh, India. His research interests include artificial intelligence, machine learning, and intelligent information systems.