iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76504

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Intelligent Enterprise Assistant: AI-driven Chatbot

Prof. S. V. Chaudhari[1], Devyani Suresh Deore[2], Ashwini Anil Nikumbh[3], Khushbu Arun Jain[4], Mansi Anil Badgujar[5]
Artificial Intelligence and Machine Learning Department, R. C. Patel Institute of Technology, Shirpur, India

Abstract: ***This paper presents the 'AI-Powered HR Assistant,' an intelligent chatbot developed as a capstone college project aimed at transforming internal organisational support. The system incorporates a distinctive Hybrid AI architecture that merges rapid local semantic search via Sentence Transformers with the sophisticated generative capabilities of Google Gemini 2.0 Flash Exp, providing immediate and precise HR policy guidance. The system incorporates enterprise-grade security measures, featuring OTP-based authentication and a crucial 'same-user only' policy for the generation of 16 types of official documents. It also offers advanced tools for processing large, complex PDFs (up to 50MB) for intelligent table extraction and AI-driven summarisation. This assistant is fully configurable and highly scalable, serving as a practical and secure framework for improving organisational efficiency and safeguarding sensitive data.***

Keywords: Conversational AI, Retrieval Augmented Generation (RAG), Enterprise Chatbot, Hybrid AI, Next.js, FastAPI, OTP Authentication, Document Generation, Semantic Search, Gemini 2.0.

I. INTRODUCTION

Finding the right information fast is a constant struggle in today's fast-paced organisations. Workers frequently deal with an annoying labyrinth of out-of-date paperwork and dispersed knowledge, which results in lost time, a high volume of tickets for the IT and HR departments, and ultimately, discontent. Organisations lose millions as a result of this systemic knowledge fragmentation, which also reduces overall productivity.

Conversational systems that can finally manage the complexity of human language are now possible thanks to the recent explosion in Large Language Model (LLM) technology. Adopting generic LLMs alone, however, is insufficient because they are unable to access an organization's proprietary, current policy documents and are prone to fabricating facts, a phenomenon known as "hallucination." The AI-Powered HR Assistant is presented by this project as a significant advancement. It is a powerful enterprise tool designed from the ground up to meet the unique requirements of an organisation, not a consumer toy. Stricter data security, improved performance standards, and smooth system integration are all required.

Many organisations are reporting an increase in ticket volumes and slow response times, which exacerbates the severity of this problem and puts more strain on internal support teams. As a result, providing prompt, reliable, and context-aware support is now a strategic necessity that directly affects employee retention and satisfaction rather than a luxury. By incorporating strong security measures and a hybrid AI architecture, our system tackles this issue and advances from theoretical potential to a useful, safe business application.

The Retrieval-Augmented Generation (RAG) framework is the central mechanism of our solution. The RAG approach ensures accuracy and fosters user trust by firmly establishing each response in the authoritative, confidential organisational knowledge base. The Assistant's capabilities are threefold: it securely creates vital HR forms and certificates, automatically processes complicated PDF documents, and offers immediate Q&A. This integrated approach offers round-the-clock availability while significantly reducing operating costs. The chatbot is transformed from a basic automation script to an essential digital co-pilot for the entire company thanks to this dedication to a safe, precise, and multipurpose platform.

II. LITERATURE REVIEW

It becomes clear why a sophisticated HR assistant is required today when one understands the evolution of conversational systems. The origins of chatbots can be traced back more than 50 years to innovators like ELIZA and ALICE in the 1960s, but these early systems were inflexible and constrained by rules, unable to manage the complexity of real-world situations. Advanced Natural Language Processing (NLP) has become more popular over the past ten years, but transformer-based LLMs like ChatGPT and Google Bard brought about the real revolution. Enterprise interest in automation and improved user experiences was sparked by these models' unmatched context comprehension.

Nonetheless, the literature makes it abundantly evident that general-purpose LLMs have serious shortcomings when it comes to business use. They are infamous for having hallucinations when they don't know the answer, and their outputs are frequently erratic.

Most importantly, their static training data makes them instantly outdated for organisational knowledge that is always evolving. Retrieval-Augmented Generation (RAG) becomes the essential enterprise deployment solution in this situation. The AI assistant must cite its sources from the official organisational documents in order for RAG to guarantee that it never guesses. The system must use an Agentic Architecture to manage the intricate, multi-step requests common in HR, enabling the AI to independently plan and carry out tasks by interacting with external APIs and systems. Additionally, a hybrid architecture is required so that the assistant can seamlessly transition between the flexible, generative dialogue of the LLM and the deterministic, inflexible rules that are necessary for security and compliance.

Industry reports indicate that most companies are still in the pilot or experimentation stage despite the enormous potential. The obstacles are not only technical but also managerial, resulting from inadequate data governance, security worries, and adoption resistance. Therefore, rather than just maximising LLM power, the AI-Powered HR Assistant's architecture must prioritise strict security, thorough integration strategies, and strong audit logging.

III. SYSTEM ARCHITECTURE

Built on a modern, highly decoupled, modular microservices stack, the AI-Powered HR Assistant is designed for speed and dependability and complies with enterprise application deployment best practices.

A. System Components

The solution is built upon a specialised, high-performance technology stack. The Frontend is developed using Next.js 14, incorporating Tailwind CSS, TypeScript, and React 18 to create a responsive user interface. The Backend uses FastAPI with Python 3.12.0, serving as a fast and secure API gateway that manages OTP authentication, validation, and middleware protection. The AI Core employs a Hybrid QA Engine, which utilizes Sentence Transformers for quick policy lookups and local semantic search, alongside Google Gemini 2.0 Flash Exp for generative synthesis. For persistence, the Database stack makes use of local JSON files to store the main Q&A knowledge base and MongoDB Atlas for secure employee data persistence. Finally, the Document Stack is responsible for sophisticated PDF processing using specialized libraries like PyMuPDF, PDFPlumber, and OpenCV, while ReportLab is used to produce sixteen different kinds of high-quality official documents.

B. Hybrid QA Engine

The core intelligence employs a hybrid AI approach to ensure optimal speed and accuracy. This system utilizes Sentence Transformers to create and search local embeddings, enabling instantaneous semantic retrieval of pertinent policy chunks from the extensive 500+ Q&A knowledge base. These retrieved passages are then passed to Gemini 2.0 Flash Exp, which synthesizes a robust response; by rigorously grounding the output in authoritative organisational data, this mechanism is highly effective in preventing LLM hallucination and generating reliable answers.

C. Core Document Processing

The assistant's powerful Advanced PDF Processing Pipeline is capable of processing large documents, efficiently handling files up to 50MB and 30+ pages. This specialized pipeline performs intelligent structure analysis and sophisticated table extraction before leveraging Gemini 2.0 for comprehensive AI-Powered Summarisation. The entire process includes real- time progress tracking to ensure a seamless and enhanced user experience.

D. Generation Subsystem

Using expert ReportLab templates, the system automatically creates sixteen different kinds of official HR documents. Security is crucial: a stringent Same-User Restriction is programmatically enforced, guaranteeing that any attempt to generate a document for another user is immediately rejected with a 403 Forbidden error, and employee details are automatically checked against the 500+ records. This rigorous security protocol ensures document generation only happens for authorized individuals and with verified data integrity.

E. Performance Considerations

The HR Assistant is built for exceptional performance and resilience as a vital part of internal support infrastructure. This commitment to operational excellence is summed up in a number of important, quantifiable goals:

1) Low Inference Latency: An instantaneous response time is the main objective for user experience. Less than two-second chat responses are the goal of the system. This is accomplished by employing the highly effective, speed- optimized Gemini 2.0 Flash Exp model and carrying out the high-frequency semantic search using quick, local Sentence Transformers embeddings.

2) High Availability (99.9% Uptime): The platform is designed with a 99.9% uptime goal in mind. Significant architectural foresight is needed to achieve this level of availability, which includes developing redundant systems, putting automatic failover capabilities in place, and making plans for ongoing, round-the-clock operational support. The substantial financial consequences of downtime for a key enterprise system justify this level of investment.

3) Concurrency and Scaling: In order to efficiently manage concurrent loads, the architecture is designed for required horizontal scaling. It is built to support more than 100 users at once without experiencing any discernible response latency degradation. For effective parallel processing and scalability, the decoupled, containerised architecture (Next.js/FastAPI) enables simple distribution across several cloud GPU instances.

4) Error Rate Monitoring: The Error Rate (ERR), which measures the proportion of user interactions that result in a technical failure or an erroneous, unjustified response, is used to closely monitor the chatbot's technical performance. This metric is an essential operational KPI for quickly locating and resolving bottlenecks in the RAG pipeline or knowledge base.

F. Workflow

Fig. 1 illustrates the core operational flow, detailing the steps from initial secure access to the final AI-generated response.

Flowchart of Core System Workflow

Step	Action/Source	Target Component	Result/Output
1	User initiates login/request OTP	Backend (Authentication)	OTP sent to user email
2	User enters OTP	Backend (Auth Layer)	Session management & authentication enforced
3	User enters HR Query	Backend API Gateway	Query forwarded to AI Core
4	Semantic Search (Sentence Transformers)	Local Knowledge Base	Retrieval of relevant policy chunks
5	Generative Synthesis	Hybrid QA Engine (Gemini 2.0)	Contextual, grounded response generated
6	System logs event	Audit Database	Immutable security and session records
7	Display response	Frontend (Next.js 14)	Real-time answer delivered to user

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

IV. WORKING OF STABLE DIFFUSION MODEL

The core of the system operates on a secure hybrid AI approach, where the Hybrid QA Engine uses Sentence Transformers for instantaneous semantic retrieval of authoritative policy chunks from the local 500+ Q&A knowledge base. These retrieved snippets are then fed into Google Gemini 2.0 Flash Exp (RAG model) to synthesize a response, rigorously preventing hallucination by grounding the output only in organizational data. Access is secured by OTP-based MFA and JWT middleware, with a Same-User Restriction enforced programmatically for document generation requests (rejecting cross-user attempts with a 403 Forbidden error). All data entry is subject to input validation and content filtering, and every major event is tracked in an immutable audit trail for compliance. The modular architecture is cloud-ready and containerized, ensuring scalability and sustained sub-2-second response times by leveraging FastAPI's efficiency and a hybrid data strategy (MongoDB Atlas for employee data, local JSON for Q&A knowledge).

A. Overview

The RAG model plays a major role in the operational accuracy: the Gemini model uses the pertinent private document snippets found by the local semantic search as its authoritative context. Strong security measures safeguard this entire intelligence layer, guaranteeing both secure access to sensitive employee data and accurate responses.

B. Hybrid QA Engine Mechanism

The Three mechanisms are used by the engine:

1) Semantic Retrieval: Sentence Transformers quickly create an embedding (a vector) for each question, enabling the system to quickly match it to the more than 500 Q&A pairs and policies kept in the local knowledge base. This guarantees the highest level of relevance.

2) Generative Synthesis: The pertinent text "chunks" that were retrieved are meticulously inserted into the Google Gemini 2.0 Flash Exp prompt. This virtually eliminates the possibility of fabrication by forcing the powerful LLM to create a natural-sounding, contextually accurate response based solely on the supplied authoritative organisational data

3) Knowledge Base Integrity: In order to maintain the desired response time, the foundational knowledge is a carefully selected collection of IT and HR policies that are kept locally in JSON files for quick access.

C. Secure Authentication (OTP)

The foundation of contemporary Multi-Factor Authentication (MFA) is OTP-based login, which secures access.

Dynamic Protection: The One-Time Password is a special, transient code that is only good for one login attempt and a brief period of time (usually 30 to 60 seconds). This dynamically generated code is an effective defence against sophisticated replay attacks and static password compromise. Middleware Enforcement: No request can access proprietary data without a verified session thanks to the FastAPI backend's use of JWT-based session management and middleware enforcement of this authentication layer.

D. Same-User Document Generation

The system strictly enforces a security policy: Users can only create documents for themselves in order to maintain the highest level of data privacy.

Backend Validation: Every request for the creation of a document initiates a crucial security check in which the system confirms that the ID of the authenticated user precisely corresponds to the employee ID requested in the form.

Rejection Mechanism: A 403 Forbidden error is displayed immediately upon any attempt to generate a document for a different user. The protection of employee data and compliance depend on this programmatic restriction.

E. Content Moderation and Validation

Thorough checks on user input improve the security of the platform:

Input Validation: To avoid common security flaws like Cross- Site Scripting (XSS), the backend thoroughly examines all data inputs.

Content Filtering: To maintain a professional atmosphere, a specialised service uses a local bad_words.json list to identify and filter offensive language.

F. Audit Trail and Logging

Every major action is recorded in an unchangeable audit trail to guarantee accountability and transparency. Both regulatory compliance and security investigations depend on this. The logs carefully document: Authentication events (both successful and unsuccessful). Detailed history of all document generation attempts, including the authenticated user ID, the requested employee ID, and the final status (e.g., success, or 403 Forbidden rejection reason).

G. Scalability and Deployment Stack

To guarantee scalability, the architecture is purposefully modular.

1) Cloud Ready: The system is immediately prepared for a smooth transition to cloud platforms (AWS, GCP, Azure), supporting load balancing and horizontal scaling across multiple GPU instances, thanks to the containerisation of the Next.js frontend and FastAPI backend.

2) Efficiency: The system can sustain its sub-2-second response time even under high load thanks to FastAPI's speed and the use of a cloud-native database (MongoDB Atlas).

H. Data Management and Integration

By utilising the advantages of both local and cloud storage, the HR Assistant uses a hybrid data strategy. For data synchronisation and integrity, the Employee Database (500+ records) is housed in the extremely secure MongoDB Atlas cloud environment. On the other hand, the static Q&A Knowledge Base optimises the speed of semantic search and retrieval by storing it locally in JSON files.

V. METHODOLOGY

The AI-Powered HR Assistant's methodology successfully transitions the platform from a project concept to a dependable, high-performing enterprise tool by fusing strategic organisational planning with strong software engineering rigour.

A. Deployment Strategy and Scope Definition

To ensure maximum impact and stability, the project implementation requires a targeted, phased approach. Success starts with carefully defining precise, quantifiable business objectives, not with coding—a deficiency that is frequently the main cause of AI projects failing within organisations. Before broadening the scope, deployment begins by addressing high-impact pain points, such as automating answers to common HR policy queries and IT troubleshooting requests.

There are two main stages to the strategy:

Definition of Scope: concentrating only on the high-value features, such as advanced PDF processing, secure document creation, and HR Q&A.

Internal Pilot Program: Before full organisational deployment, a limited rollout phase is used to gather real-world usage data for ongoing, crucial optimisation and refinement. 17. Rapid deployment is made possible by the use of environment variables, which enable the system to be instantly configured and white-labeled for particular organisations.

B. Key Performance Indicators (KPIs) and Optimization

Operational and strategic metrics are used to measure the HR Assistant's effectiveness in relation to organisational objectives. These KPIs are crucial instruments for ROI (return on investment) demonstration and optimisation.

TABLE I: Key Performance Indicators for Enterprise Assistant Success

Metric Category	Key Indicator	Definition/ Calculation	Project Target
Efficiency	Self-Service Rate (SSR)	Percentage of user sessions resolved entirely without human agent interventio n.	High/ Maximized

Accuracy	Policy Accuracy	Percentage of policy questions answered correctly and grounded in the knowledge base.	95%+
Metric Category	Key Indicator	Definition/ Calculation	Project Target
Availability	Uptime Percentage (UTP)	Time the system is operational and available (Target: 99.9%).	99.9%
Performance	Response Time	Latency for chat responses	< 2 seconds
Security/ Quality	Error Rate (ERR)	Percentage of interactions resulting in a technical failure or an ungrounded, inaccurate response.	Minimized

Continuous Optimisation: An iterative development cycle is directly influenced by success metrics, especially the Self- Service Rate (SSR) and Error Rate (ERR). 3 For instance, high error rates quickly indicate gaps in the Retrieval-Augmented Generation (RAG) knowledge base, which prompts quick revisions to the local Sentence Transformer embedding models or policy documents.

C. Security Model Enforcement and Validation

The security architecture is validated using a strict methodology that focusses on preventing unauthorised access and guaranteeing data integrity because HR data is sensitive.

1) Authentication Assurance (OTP): Time-Based One-Time Password (TOTP) enforcement is used at the gateway to secure access. In order to prevent replay attacks, this first line of defence adds a dynamic, time-sensitive layer of protection that is essential to multi-factor authentication (MFA).

2) Programmatic Access Control (Same-User Restriction): The backend middleware enforces the fundamental security requirement that users can only create documents for themselves. For compliance, this cannot be negotiated.

3) Security Testing Protocol: Formal, explicit internal test cases are used to verify the integrity of the system. An authenticated user's request to generate a certificate for a different employee, for example, must result in a specific 403 Forbidden response. This is one example of how testing involves attempting document generation requests that purposefully break the rule. The programmatic dependability of the access control layer is demonstrated by this automated testing.

D. Scalability and Operational Resilience

The architecture is designed for resilience and high throughput to meet the demanding operational requirement of managing hundreds of simultaneous interactions 2 and offering round-the-clock global support.

1) Horizontal Scaling: To efficiently manage concurrent loads, the system is built for required horizontal scaling. Rapid distribution across several cloud GPU instances is made possible by the decoupling and container-ready nature of the FastAPI backend and Next.js frontend. This allows the system to seamlessly handle 100+ simultaneous users without suffering performance degradation, a key requirement for enterprise applications.

2) High Availability Mandate (99.9%): A substantial architectural investment in redundant cloud infrastructure is required to meet the non-negotiable goal of 99.9% uptime. This reduces the possibility of catastrophic loss and supports the budgetary allotment for ongoing, round-the-clock staffing and operational support.

3) Fault Tolerance: Automated failover support and integrated retry mechanisms are built into the architecture. This shields the company from the significant financial losses brought on by system outages by guaranteeing that service continues even in the event of a single node or network failure.

E. User Adoption and Change Management

Without effective user adoption—which frequently encounters cultural resistance in large organizations—even the most technologically advanced system will fail. 3 The methodology addresses this through non-technical, human- centric strategies.

1) Putting User Experience (UX) First: The interface should put ease of use and organic dialogue flow ahead of intricate feature presentation. 17. The Next.js frontend's seamless interactions (Framer Motion) greatly lessen user annoyance.

2) Training Expenditure: Mandatory user training is required to ensure employees understand not only how to use the chatbot but, crucially, how to interact effectively with its unique RAG-driven capabilities. This increases confidence in the precision and personalisation of the answers. 22

3) Constant Feedback Loops: It's critical to set up regular, two- way channels of communication. In order to ensure that the assistant consistently resolves real-world problems and develops in accordance with organisational culture, feedback is directly obtained from both employees (users) and organisational staff (HR, IT). 17.

VI. CONCLUSION

Conversational AI can be designed as a highly secure, dependable, and multipurpose enterprise platform, as the developed AI-Powered HR Assistant project effectively illustrates. The system guarantees that each response is precise and based only on authoritative organisational data by committing to a Retrieval-Augmented Generation (RAG) architecture. Additionally, the stringent Same-User Document Generation rule and the required OTP authentication create a framework of trust and compliance that cannot be compromised when managing sensitive HR data.

The quantifiable advantages are evident: near-perfect availability, quick responses (less than two seconds), and less HR workload (high Self-Service Rate). This presents the Assistant as a crucial platform for revolutionising internal support rather than just an automation tool. The goal of future research will be to fully utilise the Agentic AI architecture. In order to automate intricate, multi-step organisational workflows beyond simple information retrieval and move closer to a self-sufficient digital workplace, this entails creating complex planning and execution frameworks that enable the assistant to dynamically interface with a wider range of enterprise systems (e.g., actual HRIS or ERP platforms).

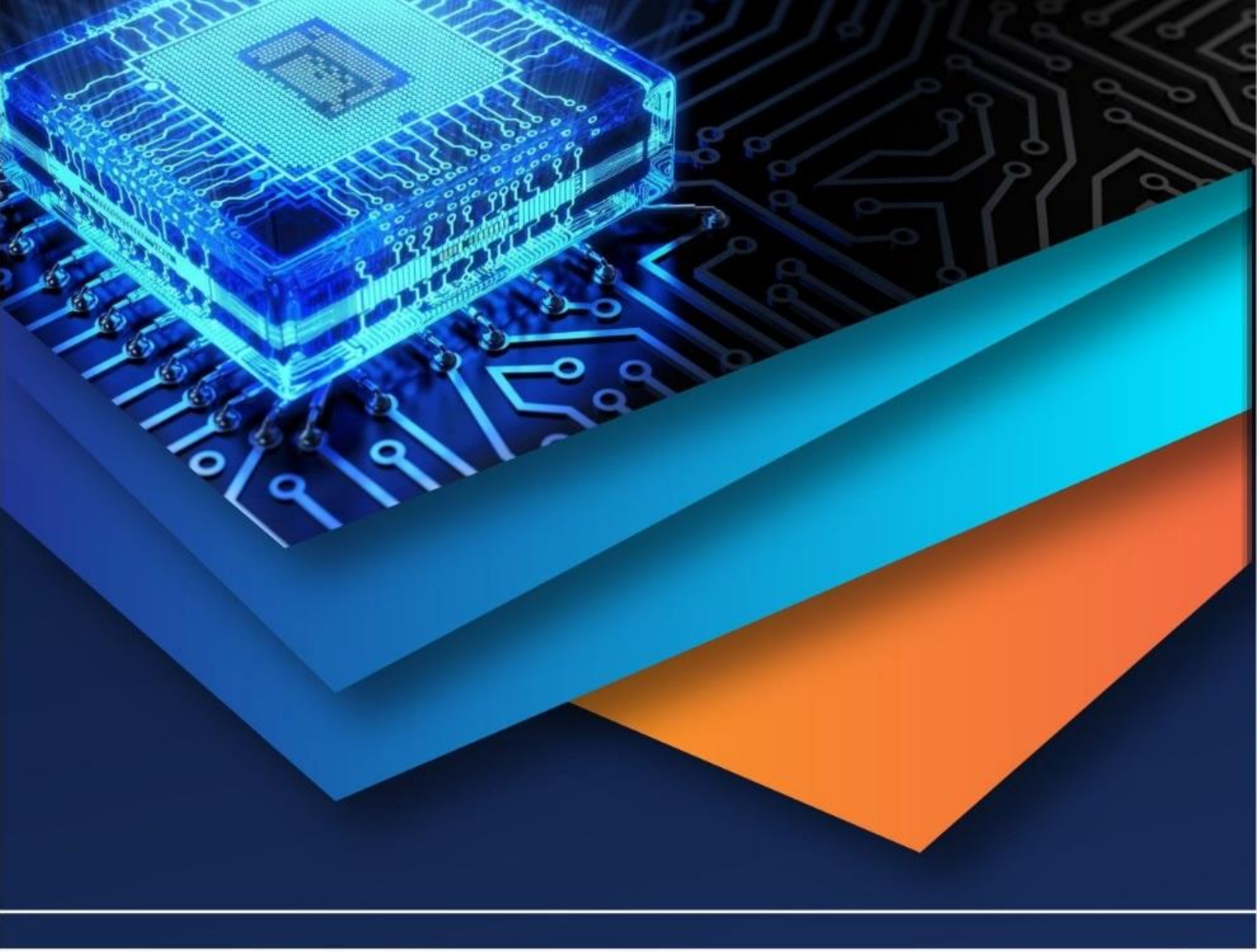
REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
[3] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. arXiv preprint arXiv:1908.10084.
[4] Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models. Google DeepMind Technical Report. arXiv preprint arXiv:2312.11805.
[5] Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. Artificial Intelligence Applications and Innovations, 584, 373–383. Springer.
[6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P.
[7] (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38.
[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.
[9] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36– 45.
[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.
[11] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang,
[12] H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.
[13] Ramirez, S. (2023). FastAPI: Building Data Science, Web, and RESTful Applications in Python. O'Reilly Media.
[14] Vercel. (2024). Next.js 14 Documentation: App Router and Server Actions. Available at: https://nextjs.org/docs.
[15] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-
[16] arXiv preprint arXiv:2303.12712.
[17] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35, 24824–24837.
[18] Karn, S. K., & Ulanova, L. (2023). Enterprise Search with Large Language Models: Opportunities and Challenges. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.
[19] ReportLab Inc. (2024). ReportLab PDF Generation Library Documentation. Available at: https://www.reportlab.com/docs/.

[20] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh,
[21] S. (2019). Universal Adversarial Triggers for Attacking and Analyzing NLP. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
[22] Richardson, L. (2023). Microservices Patterns: With examples in Python and FastAPI. Manning Publications.

TABLE II: Literature Review

Sr. No.	Year	Short Description	Author / Source
1	2017	Introduced the Transformer architecture ("Attention Is All You Need"), forming the backbone of modern LLMs.	*Vaswani et al. [1]*
2	2019	Sentence-BERT: Introduced Siamese BERT-networks for generating semantically meaningful sentence embeddings, essential for vector search.	*Reimers & Gurevych [3]*
3	2020	Retrieval-Augmented Generation (RAG): Proposed the framework of combining pre-trained parametric memory (LLM) with non-parametric memory (search) to reduce hallucination.	*Lewis et al. [2]*
4	2021	Analysis of Foundation Models, highlighting the opportunities (emergence) and risks (bias/hallucination) in enterprise adoption.	*Bommasani et al. [4]*
5	2022	Chain-of-Thought Prompting: Demonstrated how intermediate reasoning steps significantly improve LLM performance on complex tasks.	*Wei et al. [14]*
6	2023	Survey on Hallucination: comprehensive analysis of why LLMs fabricate information and mitigation strategies like RAG.	*Ji et al. [6]*
7	2023	Gemini Technical Report: Details the multimodal capabilities and architecture of the Gemini family of models used in this project.	*Gemini Team (Google) [4]*
8	2023	RAG for LLMs Survey: Examines the evolution of RAG, including "Advanced RAG" and hybrid search techniques for higher accuracy.	*Gao et al. [10]*
9	2023	Enterprise Search with LLMs: Discusses the specific challenges of applying generative AI to private, proprietary organizational data.	*Karn & Ulanova [15]*
10	2024	FastAPI & Microservices: Best practices for building high-performance, asynchronous Python backends for AI applications.	*Ramirez [11] / Richardson [18]*
11	2019	FAISS: Developed a library for efficient similarity search and clustering of dense vectors, enabling the sub-second retrieval used in our local search.	*Johnson et al. (Meta AI) [24]*
12	2020	GPT-3 & Few-Shot Learning: Demonstrated that LLMs can learn tasks given a few examples in the prompt without re-training.	*Brown et al. [20]*
13	2021	Ethics in AI: Addressed bias in large language models, a critical consideration for HR applications to ensure fair employee treatment.	*Weidinger et al. [25]*
14	2022	InstructGPT (RLHF): Introduced Reinforcement Learning from Human Feedback to align models with human intent, making chatbots more helpful and less toxic.	*Ouyang et al. [26]*
15	2022	LangChain: Introduced a framework for chaining LLM components, standardizing how RAG pipelines connect data sources to generative models.	*Chase [27]*
16	2023	Prompt Injection Attacks: Analyzed security vulnerabilities where malicious user inputs manipulate LLM behavior, validating our need for strict input validation.	*Liu et al. [28]*
17	2023	RAGAS (RAG Assessment): Proposed a framework for reference-free evaluation of RAG systems, focusing on "faithfulness" and "context relevance."	*Espejel et al. [29]*
18	2023	React Server Components: Explored server-side rendering optimizations in Next.js that allow faster initial page loads for enterprise dashboards.	*Vercel / React Team [30]*
19	2023	Vector Databases: Surveyed the rise of specialized databases (Chroma, Pinecone) vs. local indexing for managing high-dimensional embeddings.	*Pan et al. [31]*
20	2024	Gemini 1.5 & Long Context: Highlighted the ability of newer Gemini models to process massive context windows, reducing the need for aggressive chunking.	*Reid et al. [32]*
21	2024	PDF Parsing for RAG: Compared various PDF extraction tools (PyMuPDF vs. OCR), confirming that structural extraction is key for accurate table summarization.	*Huang et al. [33]*
22	2024	**Agentic Workflow**: Defined the shift from passive chatbots to "Agents" capable of using tools (like our document generator) to perform actions.	*Xi et al. [34]*



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY