SMART INDIA HACKATHON 2024

SMART INDIA HACKATHON 2024

- **Problem Statement ID –** SIH1706
- **Problem Statement Title-**
  Enterprise Assistant:
  Enhancing Organizational Efficiency
  through AI-driven Chatbot Integration
- **Theme-** Miscellaneous
- **PS Category-** Software
- **Team ID-** 1057
- **Team Name -**TECH ENERZAL

TECH ENERZAL

ENTERPRISE ASSISTANT

SMART INDIA HACKATHON 2024

Project Overview : AI-powered chatbot for optimizing organizational processes.

Cost-Effective Solution:

- Pricing is up to 30% lower than major providers.
- Fully open-source tech stack with no reliance on third-party APIs.
- Seamless integration into existing tech stack.
- Minimal oversight required for data maintenance and RAG pipeline, with synthetic FAQs dataset once generated for model training and reused for website FAQs.

Key Features:

- Automated pre-filled IT tickets based on conversation after troubleshooting.
- Automated scraping for company events data from Internal Documents/websites and Socials(LinkedIN),
- Integration with employee dashboard for granting secure access to internal documents and real-time data for the chatbot through user authentication.

Technologies & Infrastructure:

- Utilizes NLP with RAG for accurate, hallucination-free responses.
- Fine-tuned multi-agent AI architecture (combination of SLM & LLM as required).
- Achieves response time under 5 seconds with support for simultaneous users using batched responses with built in language filter

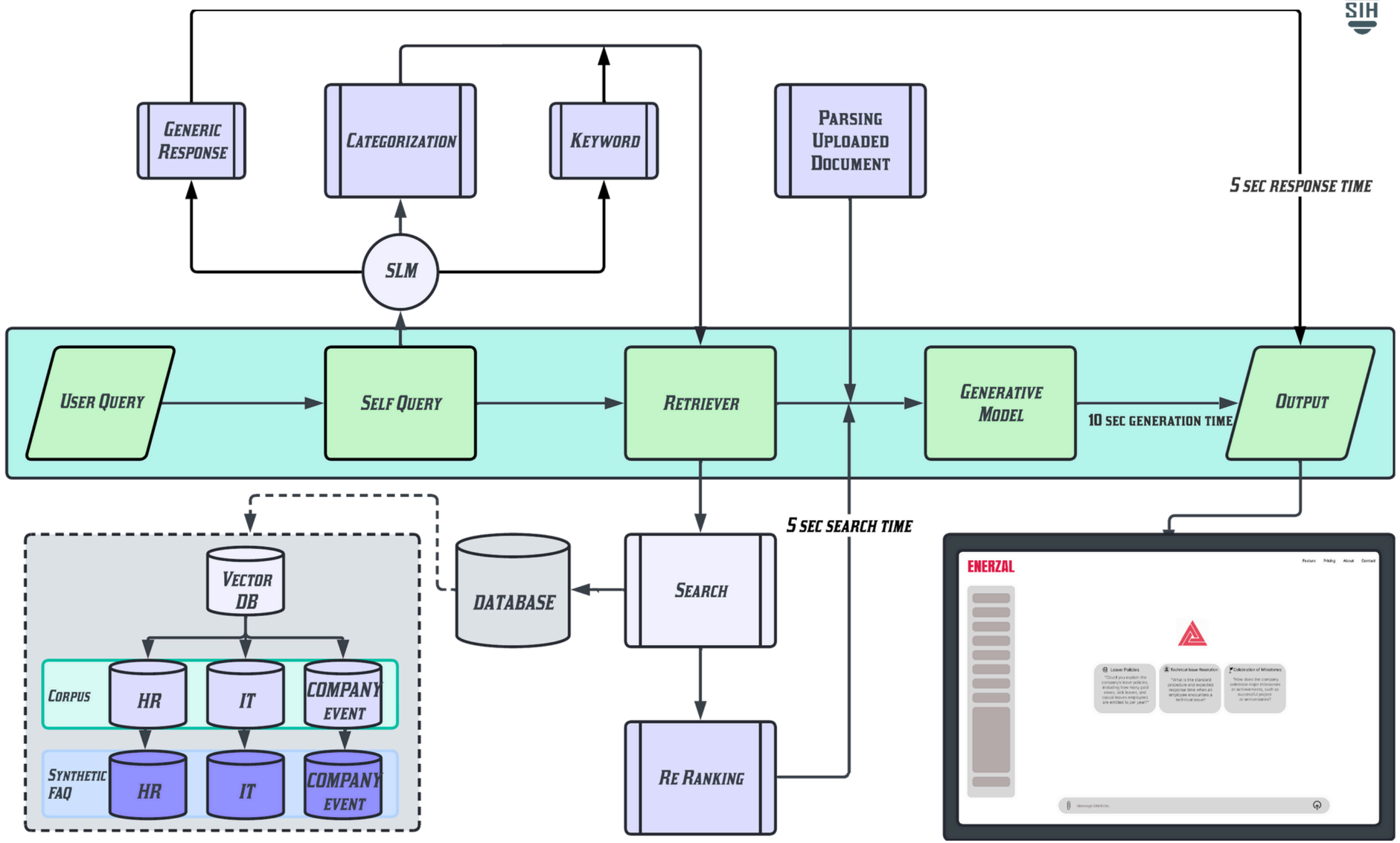
Deployment & Management:

- Data remains secure within the company's private cloud when air-gapped on local GPU infrastructure.
- Deployed using the LangChain framework with Option for serverless cloud inference for use cases not requiring air-gapped infrastructure.
- Vector database for efficient embedding storage and retrieval.

TECH ENERZAL

TECHNICAL APPROACH

SMART INDIA HACKATHON 2024



TECH ENERZAL

SMART INDIA HACKATHON 2024

FEASIBILITY AND VIABILITY

Potential challenges

- The potential risk of sensitive internal documents being exposed or leaked during data processing or system integration, especially when handling confidential information, presents a security challenge for the organization.
- Staying updated with emerging technologies from major providers is crucial to maintaining cost-effectiveness and performance advantages.
- Data maintenance for up-to-date organizational information requires constant oversight, even with automated web scraping from LinkedIn, websites, and internal documents.

Strategies for overcoming challenges

- Regularly adopt emerging technologies to enhance chatbot functionality and maintain competitive advantages.
- Optimize model efficiency and automate data processes with Graph RAG for scraped data maintenance.
- Invest in ongoing team training and leverage employee feedback for system improvements.
- Ensure data security with email/TOTP-based 2FA, database encryption, and air-gapped environments.

TECH ENERZAL

SMART INDIA HACKATHON 2024

IMPACT AND BENEFITS

Business Points:

- Increased efficiency through streamlined operations and reduced manual tasks.
- Cost savings by automating repetitive tasks and reducing labor costs.
- Improved decision-making with real-time data insights.
- Enhanced employee experience with faster, more accurate services.
- Better resource allocation for strategic, high-value activities.
- Higher employee engagement through efficient tools and streamlined processes.

Technical Points:

- Secure employee-uploaded document processing for summarization and key information extraction.
- Faster turnaround through automation of time-consuming tasks.
- Scalable to handle increased workloads without extra manpower.
- Ensures data accuracy by maintaining integrity and consistency, minimizing errors in support tasks for company events and HR policies.
- Robust encryption and security protocols for sensitive data handling.
- Seamless integration with existing systems, ensuring smooth data flow without disruptions.

TECH ENERZAL

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

RETRIEVAL AUGMENTED GENERATION: Medium , Nvidia , Google Cloud , IBM

- Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
- Salemi, Alireza, and Hamed Zamani. "Evaluating retrieval quality in retrieval-augmented generation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

RAG Framework & Pipeline: Langchain , Haystack

Vector DB: Medium , Cloudfare , Chroma DB

- Han, Yikun, Chunjiang Liu, and Pengfei Wang. "A comprehensive survey on vector database: Storage and retrieval technique, challenge." arXiv preprint arXiv:2310.11703 (2023).
- Taipalus, Toni. "Vector database management systems: Fundamental concepts, use-cases, and current challenges." Cognitive Systems Research 85 (2024): 101216.

Small Language Models: Medium , Tiny Lama

GPU Interference: Runpod.io , Genesis Cloud

2FA(TOTP/OTP): Hypr

- Seta, Henki, Theresia Wati, and Ilham Cahya Kusuma. "Implement time based one time password and secure hash algorithm 1 for security of website login authentication." 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). IEEE, 2019.