2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)
14-16 May, 2025, Astana, Kazakhstan

2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) | 979-8-3315-1514-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/SIST61657.2025.11139171

Cloistered Knowledge Capture and Retrieval: Offline LLMs and Vector Search for Enterprise

Stephen J. Hall
Department of Computer Science
Middlesex University
London, United Kingdom
SH1436@live.mdx.ac.uk
ORCID 0000-0002-6144-679X

Serengul Smith
Department of Computer Science
Middlesex University
London, United Kingdom

Can Başkent
Department of Computer Science
Middlesex University
London, United Kingdom

Clifford De Raffaele
Department of Computer Science
Middlesex University
London, United Kingdom

***Abstract* — In an era of rapid technological advancements, Artificial Intelligence (AI) has proven to be an invaluable tool for enhancing various aspects of human life. This research paper delves into the potential impact of cutting-edge AI technologies on business by integrating them into a knowledge management system. The aim of this study is to explore how these innovations can contribute significantly to empowering businesses, facilitating their growth, bolstering their competitiveness in the marketplace, and provide resilience against staff mobility. This paper examines recent advancements in Generative Transformers, Vector Databases, and Large Language Models to support the development of a novel, secure system for the capture, storage, and retrieval of knowledge—purpose-built for business-focused knowledge management. It implements and evaluates a system that challenges a sanitized version of an actual business dataset consisting of 595 documents to a set of 40 actual business enquiries. The accuracy, trustworthiness and response time of the system is evaluated with a thorough review of its limitations and recommendations for future work.**

Keywords— Generative Transformers; Vector Databases; Knowledge Graphs; Large Language Models; Enterprise Knowledge Management Systems; AI in Business Applications;

I. INTRODUCTION

Knowledge Management (KM) brings particular benefit to business enterprises by registering a lowered incidence of repeated mistakes, enhanced personal knowledge base, improved organisational competence and a saving to costs and expenses [1], [2]. These benefits are achieved through KM via the adoption of a cyclic approach to handling knowledge such that it is managed from the point of its creation up to when its value is realised and eventually no longer required and destroyed. "Knowledge management systems (KMS) are a class of information systems that manage, store and distribute knowledge. The simplest KMS's do this with explicit knowledge; advanced ones with tacit knowledge" [3]. The growing importance and role of technology advancement in enhancing KMS is multi-faceted and motivated by several key technological developments.

Technologies such as AI and Machine Learning can be integrated into KMS to improve the efficiency and effectiveness of knowledge creation, dissemination and utilization. Natural Language Processing (NLP) can also be employed to enhance the communication between machine and human language. The integration of these technologies not only enhances the capabilities of KMS but also ensures that organisations can leverage their collective knowledge more effectively, leading to improved performance, innovation and competitive advantage [4].

Existing studies predominantly rely on online AI and NLP services, which in turn raise legitimate concerns regarding data privacy and security.

This paper aims to establish how Generative Transformers, Large Language Models and Vector Databases can enable the utility of 'offline' and therefore 'secure' KM within a business enterprise.

II. LITERATURE REVIEW

NLP, text summarization, storage and retrieval are actively researched areas of linguistics and computer science. Reference [5] presents a method that ties in with the KM cycle in that it conducts text summarization "in three principal stages: knowledge acquisition, knowledge discovery, and knowledge representation. The system produces an 'abstractive summary of a given text' using syntactic analysis, NLP techniques and the 'CYC'

inference engine. The public release of the 'Generative Pre-trained Transformer'(GPT) in November 2022 significantly influenced the evolution and adoption of AI across multiple domains. Although the number of publications discussing the application of AI within the KM field is still relatively low, a notable increase in papers is being recorded [6]. At the core of this evolution is the Large Language Model (LLM) that is trained on a substantial corpus of data and forms the basis of the natural language 'Question & Answer' chatbot. Re-training a LLM is "prohibitively expensive" [7], excluding its use in fields where the corpus of data changes regularly such as in the case of KM.

Combining the advanced natural language capabilities of the LLM and the abstractive summary generation through syntactic analyses of a given text has led to the development of 'Retrieval Augmented Generation' (RAG) paradigm [8]. In its simplest form this approach uses a vector database to store and index representations of data in a form that is optimised for similarity search using Approximate Nearest Neighbor (ANN) algorithms [7]. The result is an easily updatable corpus-based information storage and retrieval system that places minimal demands on the computational effort required. In their paper entitled "AI Tools for Knowledge Management", [9] suggests the implementation of RAG to provide appropriate answers to student questions pertinent to their university. A collection of frequently asked questions together with their respective answers (QA Pairs) were embedded into vectors. The student would ask a question which was embedded into a vector, a similarity search conducted, the result retrieved from the vector database and augmented into a natural language reply by the LLM. Despite missing underpinnings to the KM principles or cycle, the system does demonstrates the potential of the RAG paradigm for explicit knowledge capture, sharing and dissemination.

A more elaborate study suggesting tighter alignment with business enterprise KM is a LLM and RAG-based QA assistant system [10]. The system was developed using a cloud-based LLM and was evaluated using the quantitative BLEU and ROUGE metrics to measure translation accuracy and summarization quality. The data sets used were based on 96 actual business documents from the human resources and information security domains. These documents were carefully curated by expert personnel and a total of 170 QA pairs were created. A variety of LLM and embedding models were used, and results yielded high BLEU and ROUGE scores demonstrating the potential that the RAG paradigm has in enterprise settings. Despite the absence of formal underpinnings to KM principles, the study demonstrates the desirable quality and accuracy aspects that RAG can provide.

Comparing the business enterprise requirements with the constraints presented by RAG exposes areas of concern. Applying such a comparison [11] determines that RAG is not yet ready for enterprise deployments in its current form. The study finds that its Accuracy, Consistency and Explainability are questionable and therefore not suited to high-risk sectors, particularly those that are compliance-regulated such as the health, legal and financial sectors. The trustworthiness of RAG is a point of contention and emanates from the inherent 'hallucination' tendencies of LLM's [12]. Despite the novelty of RAG, research suggest several enhancements to RAG architectures that claim to improve the trustworthiness aspects of LLM's. Reference [13] propose a framework consisting of "six key dimensions" to assess the trustworthiness of LLM's used in RAG systems. Factuality; Robustness; Fairness; Transparency; Accountability; and Privacy. The evolution of RAG has been rapid and yielded significant enhancements. Portrayed by a survey of RAG architectures, [8] explains the key features and strengths of each architecture – from the simplest Naïve RAG; to Advanced; Modular; Graph; and the most current Agentic RAG representing a paradigm shift in accuracy, scalability and adaptability to real-time changes.

The choice of LLM has a significant impact on the performance of a RAG system, specifically in terms of accuracy, demands on computational resources, and the time involved in query processing, retrieval, and response generation [8], [10], [14]. LLM's are employed at specific stages of the RAG pipeline depending on the complexity of the system, and it is therefore essential to consider the intended use, known strengths and weaknesses of an LLM when selecting for each specific stage [9], [10], [15], [16]. Evaluating the performance of RAG systems presents particular challenges "due to their hybrid structure and reliance on dynamic sources" [17]. The broadness of these sources necessitates metrics for evaluating accuracy of retrieved documents within the context of the query [18]. Use of publicly available datasets facilitate the comparison of results to other equivalent studies but can be highly unbalanced and lead to unreliable analogies [15], [17], [19]. The BLEU and ROUGE metrics intended for evaluating language translation and summarization tasks have been used but present fundamental shortcomings in evaluating RAG [20] particularly "low correlation with human judgements" [21]. A method found to be effective by several studies is the creation of QA Pairs curated by field experts from a given dataset pertinent to the domain of interest [15], [16], [17], [20]. This method replaces the 'context' part of the more popular QA triplet with a question classification parameter. This is found to be better suited for domain-focused datasets. Although this approach prevents comparison of system performance with other studies it does contribute toward increased RAG optimization [15], [22].

III. METHODOLOGY

To determine the feasibility of using RAG for Knowledge Capture, Storage and Retrieval, a prototype using factual business data is developed and tested. The privacy and security aspects of the system are key amongst all other functional requirements since this ensures that all sensitive company data will remain secure, local, and within the system's secured operational boundaries throughout its life cycle. To achieve this, a 'cloistered' approach is adopted whereby the system operates exclusively 'offline' and without dependency on any online/internet resources. The 'n8n' workflow

automation tool was selected as the development platform in consideration of its ability to operate off-line and integrate LLM's, the open-source Langchain AI framework, AI agents, Vector database engines, data extraction and classification functions, and file handling requirements of the system. Furthermore, its support for a highly modular development approach facilitates the configuration and replacement of function nodes with minimal distraction from the experimentation and testing phase of the study.

A. Dataset

The dataset underpinning this research comprises a curated subset of genuine business communications sourced directly from an operational IT services enterprise. It consists of 595 files in 36 folders (121 business documents and 474 email messages). Contextually, the dataset consists of Administrative, HR, Marketing Business documents, Proposals, Invoices and Email communications between the enterprise and 10 of its customers. To ensure confidentiality, all sensitive information contained within these documents has been rigorously sanitized. This process involved replacing all personal details with fictitious information such that the relationships between the communicating parties and therefore its relational integrity is maintained to all extents.

B. System architecture

The system consists of two distinct workflows - the 'ingestor' - used for Knowledge Capture, and the 'enquirer' - used for Knowledge Retrieval. The two workflows operate independently of each other and are connected only via the Knowledge Base (KB) and the user. The KB is used for storage, and the user contributes new knowledge to the data set and engages in enquiry/responses. Fig. 1. illustrates the combined 'capture and 'retriever' workflows configured to handle user enquiries via the Natural Language User Interface (NLUI) and the ingestion and storage of the dataset's three document types, i.e. email messages, invoices and other business documents.

C. Knowledge Capture / Ingestor

The capture workflow consists of four sequentially staged components. The Data Set provides the collection of Business Data to the 'Data Classifier' that analyzes the content of the data and identifies it as either an email message, invoice or other text document. According to the data type identified, the 'Data Type Handler', through customizable parameters pertinent to each type, extracts metadata and filters out any unnecessary content such as that which is duplicate, already in use as metadata, markup, and formatting information. The meta data and remaining textual content are then passed through the 'Text to Vector Encoder' and stored in the KB. The vector encoder employs a natural language embedding model to extract semantic meaning and convert the text to vectors by applying three principal parameters, the text chunk size, overlap and character separator. These parameters together with the choice of the encoder model have a significant effect on the resulting usability of the vectors and are easily configurable for the purpose of this study.

D. Knowledge Retriever / Enquirer

The retriever workflow consists of five components, all of which contribute to the handling of the user query and augmented response sub-stages. The Research Agent (RA) receives the query from the user via the NLUI, retrieves the relevant results and in combination with the initial query generates the augmented reply to form the user response. The RA is supported by a 'Session Memory' database that stores the dialog between the NLUI and RA for the duration of the session. It serves the RA to first search for responses within session memory before forwarding the request onto the Knowledge Base Retriever (KBR). This reduces system latency if the context of the query has already been retrieved or is found to be contextually valid to contribute toward the enquiry or response. The behaviour of the RA is primarily conditioned by the choice of LLM and its temperature, context length and prompt parameters which have also been made easily configurable for the purpose of this study. The KBR is responsible for all interaction with the KB. When requested to do so by the RA, it will forward a 'prompt engineered request' received from the RA to the KB via the vector encoder. The decoded response is then forwarded back to the RA for augmenting into a natural language response to the user's initial query. As with the RA, the behaviour of the KBR is also primarily conditioned by the choice of LLM, its temperature, context length and prompt parameters.

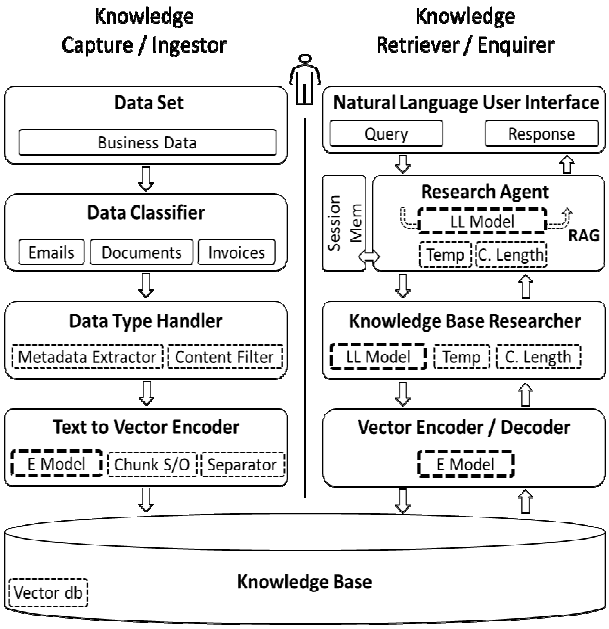

Figure 1. Knowledge Capture and Retrieval Workflows

E. User Enquiries

A set of 40 QA Pairs curated by the dataset domain experts and divided into 4 classification types (see Table 1.) will be used to evaluate the accuracy and trustworthiness of the system.

TABLE I. QA PAIR CLASSIFICATIONS

Class Type	**QA Pair Difficulty -** The Answer is:
Unanswerable	not in dataset and cannot be inferred from it.
Single Fact	in the dataset, has one unit of info, cannot be partially correct.
Summary	in the dataset, has multiple units of info, can be partially correct.
Reasoning	not explicitly in dataset but can be inferred via simple reasoning.

The 'Unanswerable' classification contains a set of 10 QA Pairs, each presenting the system with an enquiry whose answer cannot be explicitly found in, or inferred from, the KB. This set of queries is expected to produce a response such as 'information not found in the KB', e.g. "*Which part of Jupiter does ACME operate from?*". The 'Single fact' classification consists of a second set of 10 QA Pairs that presents the system with an enquiry whose answer exists explicitly in the dataset, contains a single unit of information and whose response is expected to be entirely correct without the need for inference, e.g. "*When was the NDA with MarketTravels Plc Signed?*". The 'Summary' classification consists of a further 10 QA Pairs and presents the system with an enquiry whose answer is explicitly found in the dataset but unlike the previous classification contains multiple units of information and a partially correct response is acceptable, e.g. "*Who are the signatories of the NDA between ACME and MarketTravels?*". The final set of 10 QA Pairs forms the 'Reasoning' classification and although the answer does not explicitly exist in the dataset, it should be capable of producing an acceptable response through inference, e.g. "*Propose an appropriate strategy for developing a disaster recovery plan for St. Norton*.".

F. Measurement Metric

To avoid any possibility of evaluation bias, the definition of the measurement metric is finalized before commencement of the experiments. The evaluation approach selected is score-based and largely constructed on that used by [20]. It consists of a composite of two metrics. The first metric measures the accuracy of the response and the second measures its trustworthiness. The 'accuracy' metric (A_i) consists of five 'QA Pair Evaluation Criteria' levels (see Table 2.).

TABLE II. QA PAIR CRITERIA LEVEL - SCORE ASSIGNMENT

QA Pair Evaluation Criteria	**Assigned Value**
No relevance	0
Minimal relevance but no alignment	1
Moderate relevance but inaccurate	3
Alignment but incomplete	5
Complete accuracy and alignment	7

In view of the projected dataset size, the choice of assigned values aims to amplify variations by allowing for representation of subtle differences. Each of the 40 system-generated responses is awarded a value contingent to the criteria-level achieved. A higher value indicates better performance. The 'trustworthiness' metric (T_i) determines the level to which the system is capable of producing responses that are truthful to the KB. This is measured by counting the number of responses that contain information that is not explicitly found in, or can be inferred from, the KB.

This metric yields a Boolean value and is calculated by counting the number of instances a 'hallucinated' response (or part thereof) is recorded against each of the 40 QA Pair enquiries. Equation 1 combines the accuracy *(Ai)* and trustworthiness *(Ti)* values to return the final score *(S)* against which each system-run will be evaluated. A higher (S) value indicates better overall system performance.

$$S = \left(\frac{\sum_{i=1}^{40} A_i}{280}\right)\left(1 - \frac{\sum_{i=1}^{40} T_i}{40}\right) \times 100 \qquad (1)$$

IV. IMPLEMENTATION

The system environment for this study consisted of a 6GB Graphics Processing Unit (GPU), 12-core CPU, 64GB RAM, MS Windows 11 (23H2) with subsystem for linux enabled, 'Docker' container manager, 'n8n' workflow automation tool, 'Qdrant' vector database, and the 'ollama' local model manager. In view of its native support within the 'n8n' platform the 'Qdrant' vector database and search engine is selected for the KB. The default settings of the database engine (vector size of 768 and the Cosine distance similarity algorithm) were adopted. The dataset is embedded into the KB vector database together with metadata extracted from the documents using the established 'nomic-embed-text-v1.5' embedding model. The default 'chunk-size' of 600 with an 'overlap' of 100 tokens and a dot separator value were used as the encoding parameters for the embedding model. A variety of LLM combinations were utilized for the RA and KBR throughout the runs. The selection was constrained to the latest open-source 7/8 billion parameter models consisting of diverse model architectures. Unlike the KBR which allowed for the widest variety of models, the RA was restricted to models that provided the essential 'tool' capabilities, used for making calls to the KBR.

Pre-experiment testing was conducted to establish the parameter values for the RA and KBR models, their respective user and system prompts, the number of QA Pairs and the number of runs to execute. To minimize the possibility of hallucinations and to ground the responses in factual knowledge the 'sampling temperature' parameter was set to '0'. To work within the computational resources available to the system and provide the same context window to each of the LLM's tested, the 'Context Length' parameter of the RA and KBR models was set to 2048 tokens. Reducing this value to 512 and increasing it to 4096 made no apparent difference to the pre-experiment results, thus confirming that the value was well within the contextual requirement of the system. The system and user prompts for the RA and KBR models were adjusted during pre-testing to ensure successful handling of the enquiry, calling and retrieval functionality of the KBR and generating a response based on the findings provided by the KBR to

the RA. Despite its usefulness for providing a more 'human-like' user experience, the 'Session Memory' function was disabled during system runs as pre-testing exposed sequential enquiry/response contamination, resulting in compounded non-identical responses to identical enquiries.

Data collection was carried out over a total of 11 identical system runs, distinguished only by a different combination of models assigned to the RA and KBR. Each run consisted of a timed session where the 40 enquiries were entered sequentially, with results registered on a score card. The system-run duration was divided by the number of enquiries to yield the average response time per enquiry (D). A score was assigned to each of the responses, any instances of hallucination registered, and values for Ai, Ti and S calculated. The results of the system-runs are reflected in Table 3. and details of the models used, and scores assigned according to their classification. No score was added to a response that contained information beyond what was declared in the QA Pair expected response. Furthermore, although an essential aspect to this study, it should be noted that the system run 'Time' values do not influence S, since its purpose is to provide a comparative measure across runs and give an indication of system efficiency compared to traditional methods of enterprise knowledge retrieval.

V. Results and Discussion

The objective of this investigation is to assess the systems potential to provide benefits to the business by way of time and manpower saved in retrieving valuable enterprise knowledge effectively and efficiently. The results in Table 3. show a detailed analysis of the scores achieved by the system.

A. Response time (D)

The longest average response time per enquiry (D) recorded was 75 seconds compared to the shortest of just 15 seconds. It is worth noting that the shortest *D* did not yield the lowest overall S but neither did the longest yield the highest. In fact, the highest *S* calculated was 0.82, which took an average 69 seconds to produce a response. The lowest overall *S* of 0.46 took 26 seconds to generate a response. Although this suggests that there is no direct correlation between *D* and the quality of the response, further study is required to determine this with certainty. Nevertheless, and despite the longest recorded *D* of 75 seconds, having access to this level of knowledge in the business enterprise setting within such a short time could comparably be considered instant.

B. Accuracy and Trustworthiness

Accuracy and Trust are essential to instill system confidence in enterprise staff. In the absence of this confidence the system will fall into disuse and be considered a waste of company resources. A graphic representation of how each system run performed in terms of Accuracy, Trust and the resulting Score is shown in Fig. 2.

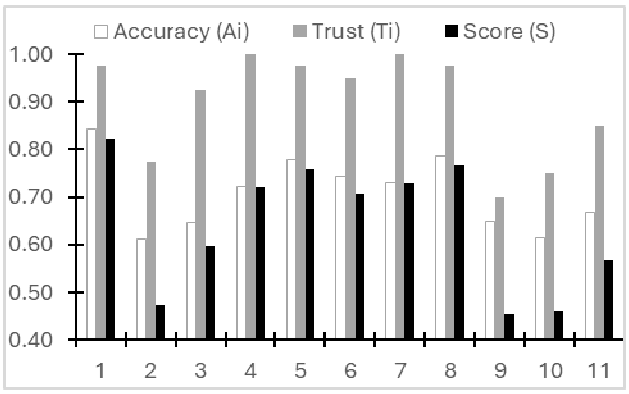

Figure 2. System Performance as a function of Accuracy and Trust

The highest recorded A of 0.84 in combination with a T of 0.98 yielded an overall system score of 0.82. In contrast, the lowest recorded A of 0.61 and T of 0.70 yielded an overall system score of 0.46.

TABLE III. Performance Results of Proposed KM Architecture on the Execution of 11 System Runs with Permutations of RA and KBR Models for QA Pair Classifications: Unanswerable (Un), Single Fact (Sf), Summary (Su) and Reasoning (Re) on Accuracy, Trust and Score.

Run	RA Model	KBR Model	Time (D)	Accuracy Un	Sf	Su	Re	(Ai)	Trust Un	Sf	Su	Re	(Ti)	Score (S)
1	Mintron 4B Instruct	Mistral 7B Instruct v0.3	69	0.94	0.83	0.71	0.89	**0.84**	0.90	1.00	1.00	1.00	**0.98**	0.82
2	Mintron 4B Instruct	Mintron 4B Instruct	15	0.60	0.61	0.63	0.60	**0.61**	0.50	0.80	0.80	0.10	**0.78**	0.47
3	llama 3.2	llama 3.2	29	0.81	0.66	0.57	0.54	**0.65**	0.90	0.10	0.90	0.90	**0.93**	0.60
4	Mintron 4B Instruct	llama 3.2	32	0.97	0.69	0.69	0.54	**0.72**	1.00	1.00	1.00	1.00	**1.00**	0.72
5	Mintron 4B Instruct	Qwen2-7B-Instruct	36	1.00	0.83	0.66	0.63	**0.78**	1.00	1.00	1.00	0.90	**0.98**	0.76
6	Mintron 4B Instruct	DeepSeek R1-7b	75	0.83	0.60	0.71	0.83	**0.74**	0.80	1.00	1.00	1.00	**0.95**	0.71
7	Mintron 4B Instruct	Granite3-dense	47	0.89	0.71	0.66	0.66	**0.73**	1.00	1.00	1.00	1.00	**1.00**	0.73
8	llama 3.2	Mistral-7B-Instruct-v0.3	44	0.89	0.74	0.69	0.83	**0.79**	0.90	1.00	1.00	1.00	**0.98**	0.77
9	Mintron 4B Instruct	llama3-chatqa	26	0.49	0.69	0.69	0.74	**0.65**	0.30	0.80	0.70	1.00	**0.70**	0.46
10	llama3.2	llama3-chatqa	39	0.49	0.77	0.54	0.66	**0.61**	0.60	0.90	0.70	0.90	**0.75**	0.46
11	llama3.2	Mintron 4B Instruct	24	0.70	0.71	0.63	0.63	**0.67**	0.60	0.90	0.90	0.10	**0.85**	0.57

The evident lack of correlation between A and T suggests the importance of having these two metrics calculated separately to determine the utility that can be expected from such a system.

C. QA Pair classifications

Further insight into the A and T aspects of the system is provided by the 4 QA Pair classifications and their direct relevance to internal business enterprise knowledge. Interestingly, neither of the 2 system runs that yielded an absolute 1.0 achieved the highest S. In fact, the run that

recorded the highest S registered a T of 0.98 showing some evidence of hallucination in the 'unanswerable' classification. This same run registered the highest A of 0.84 and was the strongest in the Single Fact, Summary and Reasoning classification, and only third highest in the Unanswerable class. Despite the variances in performance, and the broadness of accuracy and trustworthiness figures, the results suggest potential benefits that such a system could yield to the business are measurably advantageous.

VI. Conclusions

This study evaluates the potential utility that can be leveraged from the latest developments in Generative Transformers, Vector Databases and Large Language Models for knowledge management systems in corporate settings. It presents the findings of a comprehensive workflow that integrates these technologies into a knowledge capture, storage and retrieval system attuned to the business requirements of enterprise. The proposed architecture enhances the retrieval of business knowledge and expertise, provides improved business decision-making, and resilience against corporate amnesia resulting from staff mobility and unscheduled terminations. The system operates without any online dependencies ensuring the privacy necessitated by sensitive company data. A thorough system evaluation resulted in highly positive findings in terms of system accuracy, trustworthiness and time to respond, and generated valuable responses to reasoning and suggestive enquiries. Future research should consider extracting more granular metadata from the dataset and introduce an additional metric that measures the perceived value of the responses.

References

[1] S. L. Cheng and W. Kuan Yew, "Knowledge management performance measurement in micro-, small-, and medium-sized enterprises An exploratory study," *Bus. Inf. Rev.*, vol. 32, no. 4, pp. 204–211, 2015, [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/0266382115615262

[2] S. J. Hall and C. De Raffaele, "Corporate amnesia in the micro business environment," in *2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, 2013.

[3] R. V Mccarthy, J. E. Aronson, and L. H. Embry-Riddle, "Success Stories in Knowledge Management Systems," *Issues Inf. Syst.*, no. December, 2017, doi: 10.48009/1_iis_2017_64-77.

[4] S. Hall, S. Smith, C. Baskent, and C. De Raffaele, "Knowledge Management for the Micro Enterprise: A Taxonomy," *Proc. Eur. Conf. Knowl. Manag. ECKM*, vol. 1, pp. 491–498, 2023, doi: 10.34190/eckm.24.1.1268.

[5] A. Timofeyev and B. Choi, *Knowledge based system for composing sentences to summarize documents*, vol. 976, no. March. Springer International Publishing, 2019. doi: 10.1007/978-3-030-15640-4_9.

[6] H. Taherdoost and M. Madanchian, "Artificial Intelligence and Knowledge Management: Impacts, Benefits, and Implementation," *Computers*, vol. 12, no. 4, 2023, doi: 10.3390/computers12040072.

[7] K. W. Church, J. Sun, R. Yue, P. Vickers, W. Saba, and R. Chandrasekar, "Emerging trends: A gentle introduction to RAG," *Nat. Lang. Eng.*, vol. 30, no. 4, pp. 870–881, 2024, doi: 10.1017/S1351324924000044.

[8] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," 2025. [Online]. Available: http://arxiv.org/abs/2501.09136

[9] M. Pondel *et al.*, "AI Tools for Knowledge Management – Knowledge Base Creation via LLM and RAG for AI Assistant," *Lect. Notes Networks Syst.*, vol. 1218 LNNS, pp. 3–15, 2024, doi: 10.1007/978-3-031-78468-2_1.

[10] G. Şahin, K. Varol, and B. K. Pak, "LLM and RAG-Based Question Answering Assistant for Enterprise Knowledge Management," in *UBMK 2024 - Proceedings: 9th International Conference on Computer Science and Engineering*, IEEE, 2024, pp. 157–162. doi: 10.1109/UBMK63289.2024.10773564.

[11] T. Bruckhaus, "RAG Does Not Work for Enterprises," pp. 1–14, 2024, [Online]. Available: http://arxiv.org/abs/2406.04369

[12] J. Sun *et al.*, "Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph," *12th Int. Conf. Learn. Represent. ICLR 2024*, pp. 1–31, 2024.

[13] Y. Zhou *et al.*, "Trustworthiness in Retrieval-Augmented Generation Systems: A Survey," 2024. [Online]. Available: http://arxiv.org/abs/2409.10102

[14] F. Cuconasu, G. Trappolini, N. Tonellotto, and F. Silvestri, "A Tale of Trust and Accuracy: Base vs. Instruct LLMs in RAG Systems," 2024.

[15] R. T. de Lima *et al.*, "Know Your RAG: Dataset Taxonomy and Generation Strategies for Evaluating RAG Systems," 2024. [Online]. Available: http://arxiv.org/abs/2411.19710

[16] H. Emekci, "Maximizing RAG efficiency : A comparative analysis of RAG methods," *Nat. Lang. Process.*, vol. 31, no. 1, pp. 1–25, 2025, doi: 10.1017/nlp.2024.53.

[17] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," pp. 1–20, 2024, [Online]. Available: http://arxiv.org/abs/2405.07437

[18] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries," 2024, [Online]. Available: http://arxiv.org/abs/2401.15391

[19] C. Zhang, V. Datla, and A. Shrivastava, "An Automatic Method to Estimate Correctness of RAG," in *Proceedings of the 31st International Conference on Computational Linguistics*, Association for Computational Linguistics, 2024, pp. 603–611.

[20] P. Finardi *et al.*, "The Chronicles of RAG: The Retriever, the Chunk and the Generator," 2024, [Online]. Available: http://arxiv.org/abs/2401.07883

[21] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 2511–2522, 2023, doi: 10.18653/v1/2023.emnlp-main.153.

[22] I. Papadimitriou, I. Gialampoukidis, S. Vrochidis, Ioannis, and Kompatsiaris, "RAG Playground: A Framework for Systematic Evaluation of Retrieval Strategies and Prompt Engineering in RAG Systems," 2024. [Online]. Available: http://arxiv.org/abs/2412.12322