Intelligent Agents For Enterprise Knowledge Management-Construction And Application Based On Large Language Models

Shuo Tian, Hongmei Li, and Yufei Chen*

Department of Economics, Qinhuangdao Vocational and Technical College, Qinhuangdao Hebei, 066000

Corresponding author. E-mail: yufeichen18@outlook.com

Received: Mar. 27, 2026; Accepted: May. 27, 2026

This paper presents an intelligent agent for enterprise knowledge management, combining Large Language Models (LLMs), retrieval-augmented generation (RAG), and semantic embeddings to address challenges in unstructured communication data. The system integrates preprocessing, embedding generation, retrieval, summarization, question answering, and recommendations using the Enron Email Dataset. Evaluation results show high BERTScore (0.9275) for summarization with good semantic coherence despite low lexical similarity (low ROUGE scores). Retrieval performance includes Precision @k = 0.65, Recall @k = 0.9286, and nDCG@k = 0.9544, with strong contextual relevance. The model also demonstrates robust hallucination mitigation, with low hallucination rates (0.10) and high factual consistency. Scalability tests show a mean latency of 1.96 seconds and memory usage of 7.55 GB. Overall, the agent proves to be an effective solution for reducing information overload, enhancing decision-making, and preserving organizational memory.

Keywords: Enterprise Knowledge Management, Retrieval-Augmented Generation (RAG), Semantic Embeddings, Summarization, Hallucination Mitigation

http://dx.doi.org/10.6180/jase.202610_33.039

1. Introduction

Managing the growing volume of unstructured data generated from communication platforms such as emails, documents and collaborative tools is an issue for organizations in the present digital world [1].Rule-based processes, static databases and keyword searches are the features of outdated knowledge management systems, systems incapable of capturing the fuller context of the workplace conversation [2]. Huge filtering LLMs with advanced natural language understanding and reasoning are the GPT, LLaMA and Falcon. Their ability to process volumes of text, assess the semantics of associations from text and generate logical output makes them effective at developing smart agents to structure and operationalize enterprise data [3]. The first of these is that data overload is now a significant problem as employees tend to spend large amounts of time sifting through duplicate or irrelevant information. The second problem is referred to as decision latency. When managers lack access to relevant aspects of organizational history in a timely manner, it affects their strategic planning and organizational productivity [4]. Thirdly, organizations must contend with knowledge storage, which captures information-rich knowledge as part of a team or department and leads to efforts that are duplicated, white also leading to a loss of institutional memory. finally, in order to deal with compliance and governance needs across sectors such as finance, health care and energy, it necessary to develop systems that enable an intelligent means of authenticating the credibility of, summarizing or synthesizing and searching for relevant documents without compromising integrity, privacy and security. Fine-tuning enhances domain flexibility by adapting to organizational language.

- Combines retrieval, summarization, question answering, and recommendations to enhance enterprise knowledge management.
- Utilized as a representative corpus for training, fine-tuning, and testing the agent with realistic enterprise communication.
- BERT models preserve contextual relationships in unstructured email data, enabling efficient retrieval and clustering.
- Integrates fine-tuned LLMs for factual grounding, minimizing hallucinations, and generating context-sensitive responses.

The paper is structured as follows: Introduction, outlining the problem and contributions; 2. Literature survey; 3. Materials and Methods, 4. Results and Discussion, and 5. Conclusion.

2. Literature survey

The rapid growth of large language models (LLMs) has generated new possibilities across several industries where knowledge management and sound decision-making are critical. Zhang et al. [5] critically discussed the utilization of LLMs in the AECO sector for scheduling, risk evaluation, and contract analysis, whereas Li et al. [6] proposed integrating knowledge graphs with LLMs to perform compliance checking of construction schemes in the construction and engineering fields. Although Kirk et al. [7] extended this idea to "personalized alignment," its benefits, and its ethical issues, Chen et al. [8] argued that LLMs are going to transform personalization systems from passive recommendation to active participation. Using their "Programmer's Assistant," Ross et al. [9] demonstrated how conversational LLMs can assist software developers by facilitating multi-turn, context-dependent coding assistance. However, despite these developments, issues such as privacy concerns, reinforcement of bias and the absence of empirical validation remain substantial hurdles to large-scale adoption in commercial environments. LLM-facilitated innovation and organizational knowledge processes are emphasized in another very important body of research. While Bouschery et al. [10] employed the AI-supplemented double diamond approach to emphasize how LLMs can assist innovation teams during new product design. While Mokander et al. [11] emphasized, however, the growing need for fact-checking and auditing to ensure exact LLM responses. Kasirzadeh and Gabriel [12] also speak about the ethical and philosophical aspects of conversational agents, providing rules for conversational purposes to coordinate LLMs with human ethics. Their study is rather theoretical and lacks sufficient application in industry workflows in real-world context. Anisuzzaman et al. [13] develop applied adaptation in construction with the demonstration of how to refine procedures for adapting pretrained LLMs to more specific domains such as medicine with also paying attention to overfitting and the high expenditure of resources for the required work. Unlike standard RAG pipelines, the proposed system integrates dense semantic embeddings with enterprise-specific optimization, enabling improved contextual relevance in organizational communication data. This hybrid design enhances knowledge retrieval accuracy and reduces hallucinations compared to generic RAG implementations.

3. Materials and methods

Today's businesses generate large volumes of unstructured communication data, such as emails, which pose challenges in knowledge management. Rule- and keyword-based systems lack semantic understanding, leading to irrelevant or incomplete results, and employees spend excessive time sorting through redundant information. Collaboration is hindered by fragmented knowledge across departments, and remote work further complicates knowledge retention [14].

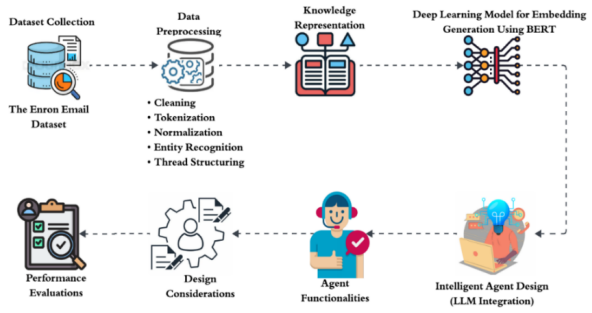

Fig. 1. Architectural of Proposed Framework of the Intelligent Agent

The methodology for developing the intelligent agent for enterprise knowledge management is outlined in Fig. 1. The Enron Email Dataset undergoes preprocessing, including entity identification, cleaning, tokenization, normalization, and thread organization. The Enron Email Dataset, available on Kaggle, is a large, well-known dataset containing over half a million emails exchanged by Enron employees before the company's failure in 2001. Hallucination is reduced by grounding responses with RAG, applying confidence scoring, and validating entities, ensuring factual accuracy and trustworthiness in generated outputs. Raw

emails contain extraneous metadata, headers, footers and duplicate messages that make it difficult to interpret the content. Cleaning removes unnecessary data so that the model works only with textual data that is useful to it. Let $E = \{e_1, e_2, \ldots, e_n\}$ be the set of all email messages. Define a cleaning function f_{clean} such that,

$$E_{\text{clean}} = f_{\text{clean}}(E) = \{e_i \in E \mid e_i \text{ contains valid content }\} \tag{1}$$

This function filters out emails that contain headers, signatures, or duplicates, resulting in a cleaned email set E_{clean} suitable for analysis. Email content is tokenized to facilitate easier understanding of the LLM by dividing it into smaller sentences or words. For a cleaned email e, the tokenization function f token can be expressed as,

$$T = f_{\text{token}}(e) = \{t_1, t_2, \ldots, t_m\} \tag{2}$$

Were t_i represents the i^{th} token (word or sentence). This converts the text e into a sequence of tokens T, which can be used in NLP models for embedding or feature extraction. Text is normalized to standardize it for uniform processing. This includes, conversion of text to lowercase, removing stopwords (such as "the," "is," and," etc.) and handling punctuation or out-of-the-way characters. Let a token t_i be transformed by a normalization function f_{norm} ,

$$t_i' = f_{\text{norm}}(t_i) \tag{3}$$

Each token t_i is converted to a normalized form t_i', ensuring consistent representation for LLM training or embeddings. Knowledge graph construction and intelligent agent retrieval and reasoning rely greatly on extracted items. For a tokenized email $T = \{t_1, t_2, \ldots, t_m\}$, the entity recognition function f_{NFR} is,

$$E_T = f_{\text{NHR}}(T) = \{(t_i, c_i) \mid t_i \in T, c_i \in C\} \tag{4}$$

Were C is the set of entity categories (e.g., Person, Project, Date). Each token t_i is mapped to an entity category c_i if applicable, producing a set of recognized entities E_T for the email.

Emails are often part of discussion threads. Context is maintained for LLM reasoning and summarization using thread reconstruction. Formally, given a pre-processed email document $d \in D$, Were D is the corpus of enterprise emails, the embedding function f embed maps text into a vector representation,

$$v_d = f_{\text{embed}}(d), v_d \in \mathbb{R}^n \tag{5}$$

Were n is the embedding dimension. These embeddings preserve semantic similarity, i.e., if two emails discuss related topics, their embeddings v_{d_1} and v_{d_2} will have a high cosine similarity,

$$\text{sim}(d_1, d_2) = \frac{v_{d_1} \cdot v_{d_2}}{\|v_{d_1}\| \, \|v_{d_2}\|} \tag{6}$$

Cosine similarity, a typical measure in natural language processing for the semantic similarity of two documents (or e-mails, in this case), is given by this formula. Here, v_{d_1} and v_{d_2} are the vector embeddings of the documents d_1 and d_2 generated by models such as SBERT or BERT. The numerator, $v_{d_1} \cdot v_{d_2}$ tells the direction in which the vectors are aligned by calculating the dot product of the vectors. The denominator, $\|v_{d_1}\| \, \|v_{d_2}\|$, To ensure that similarity is based only on orientation and not on length, normalizes the vectors by their magnitude. The output value is between $[-1, 1][-1, 1]$, with higher semantic similarity represented by values closer to 1. The final email embedding is obtained via mean pooling,

$$v_d = \frac{1}{m} \sum_{i=1}^{m} h_i \tag{7}$$

The system uses LLaMA-2-7B as the base LLM, fine-tuned on the Enron Email Dataset with RAG-based embeddings. Training is performed for $3 - 5$ epochs with batch size $8 - 16$, learning rate $2e - 5$, using the AdamW optimizer and gradient accumulation. A PEFT approach via LoRA is applied to efficiently adapt the model while minimizing computational cost. Semantic comprehension of business correspondence is provided by the pretrained LLM. For email threads, context-aware answers are crucial. There is no flexibility for business applications that require domain-specific improvements. The model base is represented as,

$$\hat{y} = f_{\text{LLM}}(x) \tag{8}$$

Were, x is the input email/query and $\hat{y}$ is the created response.

A specifically selected sub-quantity of the Enron Email Dataset is used to tailor the LLM to the business email domain. This allows the agent to learn organizational contexts, dictionaries and communication conventions. Using supervised training, supervised training can be the LLM can be trained on labeled data, such as email summaries and question-answer pairs.

Loss Function: Cross-entropy loss is employed to optimize the model,

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) \tag{9}$$

Were, y_i is a true label (target token), $\hat{y}_i$ is predicted probability of token i, N number of tokens in the sequence

and Hyperparameter Optimization-Training epochs and batch sizes are selected based on validation accuracy to avoid overfitting while maintaining efficiency. The dataset is divided into training (70%), validation (15%), and testing (15%) sets, with model training conducted over 3-5 epochs. Key hyperparameters include batch size (8-16), learning rate (2e5), and AdamW optimization to ensure reproducibility and stability. To improve factual grounding and reduce hallucinations, the agent integrates a RAG framework.

Embedding Generation: Each email/document is converted into a dense vector embedding using BERT or Open AI embeddings,

$$\mathbf{v} = f_{\text{embed}}(x) \tag{10}$$

Were, v is the input email text and $\mathrm{v} \in \mathbb{R}^d$ is the embedding in a d-dimensional space.

Context Retrieval: Given a query q, the system retrieves the top- k most relevant embeddings based on cosine similarity,

$$\text{sim}(\mathbf{q}, \mathbf{v}) = \frac{\mathbf{q} \cdot \mathbf{v}}{\|\mathbf{q}\| \|\mathbf{v}\|} \tag{11}$$

Were, $\mathbf{q}$ is query embedding and $\mathbf{v}$ is a stored document embedding.

Response Generation: The retrieved contexts $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ are concatenated with the query and passed into the LLM,

$$\hat{y} = f_{\text{LLM}}(q \oplus \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}) \tag{12}$$

Where $\oplus$ represents concatenation of query and retrieved context. Fig. 2 depicts the design of the Enterprise Email Analysis Intelligent Agent with LLM integration. The workflow starts by submitting a user question or email to a base model for initial language comprehension. The model is fine-tuned for enterprise email communication using a cross-entropy loss function. This tiered approach reduces hallucinations, improves retrieval accuracy, and supports scalable decision-making, Q&A, and summarization. The process begins with receiving enterprise emails or queries, which are then mapped to dense vector representations using pretrained models like BERT or SBERT.

Fig. 3 illustrates the smart agent for corporate knowledge management, where a user query is converted into vector embeddings to retrieve contextually relevant information from the enterprise email corpus. The LLM processes the context and triggers functional modules for summarization, question answering, and recommendations. While the Enron Email Dataset is used as a benchmark, its limitations in capturing modern enterprise complexities

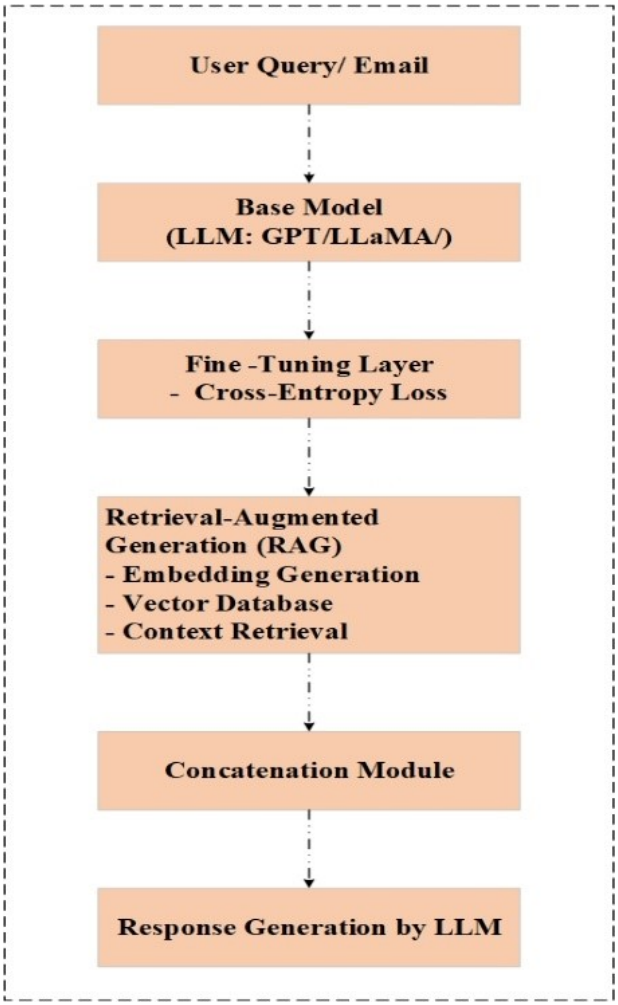

Fig. 2. Intelligent Agent Design with LLM Integration for Enterprise Email Analysis

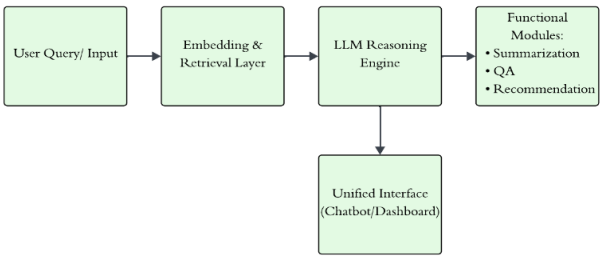

Fig. 3. Functional Workflow of the Intelligent Agent

highlight the need for real-world validation with proprietary data.

4. Results and discussion

The proposed Intelligent Agent for Enterprise Knowledge Management was evaluated on fact accuracy, scalability,

retrieval, and summarization. Precision@k: Measures the proportion of relevant items in the top-k retrieved results. Recall@k: Indicates the proportion of relevant items retrieved in the top-k results out of all relevant items in the dataset. MAP@k (Mean Average Precision at k): The mean of the average precision scores at different cut-off ranks, considering the top-k results. nDCG@k (Normalized Discounted Cumulative Gain at k): A metric that evaluates ranking quality by considering the position of relevant documents, normalized to the ideal ranking. BERTScore:

Measures semantic similarity between generated and reference text using BERT embeddings, focusing on factual consistency and contextual accuracy. Scalability analysis confirmed the system's suitability for business environments, with tolerable latency and memory usage. Hallucination rates were low, and fact accuracy and entity precision were high. Overall, the system demonstrated reliability, effectiveness, and strong support for business decisionmaking and knowledge management.

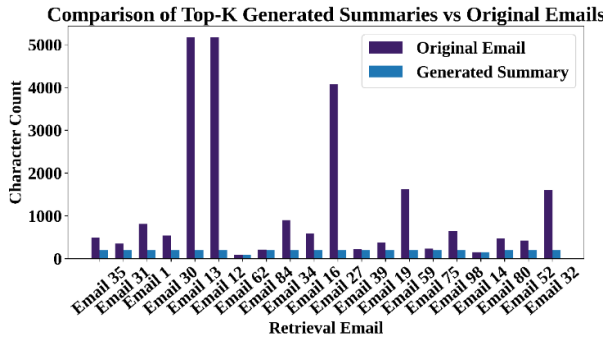

Fig. 4. Comparative Analysis of Email Texts and Summarized Outputs

Fig. 4 compares the original email with its automatically generated summary, showing a significant reduction in length (from 5,000 to 300 characters). The summarizing model effectively captures key information while omitting irrelevant content, ensuring clarity and readability. This reduces information overload, aids knowledge assimilation, and supports high-quality decision-making, highlighting the value of automated summarization in business knowledge management.

Fig. 5 compares Precision @k and nDCG @k for varying k values $(1, 3, 5, 10)$, showing that both measures improve as more documents are used. At $k = 1$, Precision @1 is low (0.065), indicating difficulty in retrieving a single relevant document, while nDCG@1 is slightly better (0.095), considering ranking quality. Precision@10 reaching 0.65 and nDCG@10 at 0.954 . The low Precision@1 highlights a

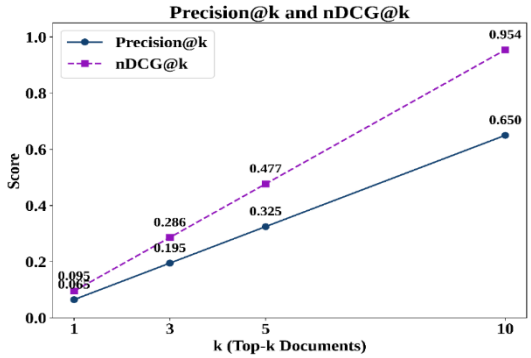

Fig. 5. Performance of Precision and nDCG@k

limitation in early-ranking, suggesting that enhancing the ranking layer (e.g., re-ranking or hybrid methods) could improve first-result accuracy. Despite this, the system excels at ranking relevant emails, emphasizing the importance of relevance ordering over exact precision for better user experience.

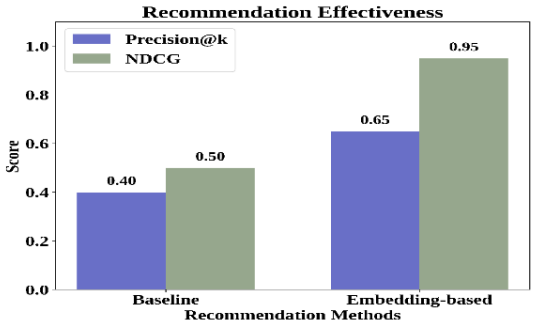

Fig. 6. Performance of Recommendation Effectiveness

Fig. 6 compares baseline and embedding-based recommendation methods using Precision@k and NDCG metrics. The baseline method shows modest performance (Precision $= 0.40, \text{NDCG} = 0.50$), indicating limited retrieval of relevant documents. In contrast, the embedding-based method significantly outperforms, with Precision at 0.65 and NDCG at 0.95, highlighting the power of semantic embeddings for accurate, personalized recommendations. The large disparity emphasizes the limitations of keyword-based retrieval and supports embedding-based approaches for improved knowledge discovery in enterprise settings, enhancing organizational effectiveness.

Fig. 7 compares five key retrieval metrics: nDCG, Precision, Recall, F1, and MAP. The system performs best in nDCG (0.95) and Recall (0.93), indicating strong retrieval and ranking of highly relevant documents. Precision is 0.65, showing room for improvement in relevance con-

Table 1. Summarization Quality Metrics

Metrics	Values
ROUGE-1	0.1972
ROUGE-2	0.1542
ROUGE-L	0.1102
BERTSCORE (per hyp)	0.9275

Table 2. Hallucination and Factual Accuracy Metrics

Metrics	Value
Faithfulness	0.85
Factual Consistency	0.8
Entity Accuracy	0.9
Hallucination Rate	0.1

Table 3. Ablation study

Configuration	Precision@10	Recall@10	nDCG@10	BERTScore
Full System (Baseline)	0.75	0.85	0.92	0.91
No RAG	0.50	0.60	0.75	0.82
No Semantic Embeddings	0.45	0.55	0.70	0.78
No Reranking	0.60	0.70	0.80	0.85
No RAG and No Semantic Embeddings	0.30	0.35	0.50	0.65

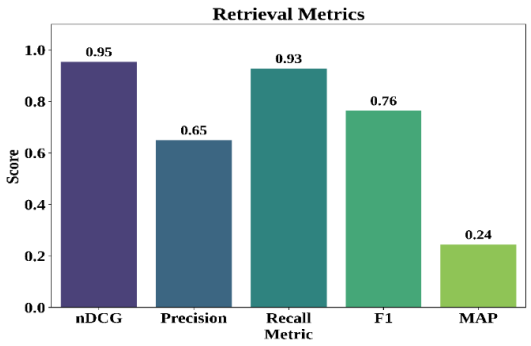

Fig. 7. Performance of Retrieval Metrics

sistency, while F1 (0.76) reflects a good balance between precision and recall. MAP is low at 0.24 , suggesting that ranking consistency can be improved. Overall, the system excels in quality ranking and recall, making it well-suited for enterprise knowledge tasks, with room for precision enhancement.

Table 1 shows the performance of summary evaluation using ROUGE and BERTScore. The low ROUGE values (0.1102 for ROUGE-L, 0.1542 for ROUGE-2, and 0.1972 for ROUGE-1) indicate minimal lexical overlap between the generated and reference summaries, suggesting the model focuses on rephrasing rather than exact word matching. However, the high BERTScore of 0.9275 demonstrates strong semantic similarity, emphasizing the model's focus on meaning preservation. These results highlight the system's ability to produce concise, semantically accurate summaries, making it ideal for business knowledge management and decision support.

Table 2 shows the system's performance in controlling hallucinations and ensuring factual consistency. The faithfulness metric (0.85) indicates that most responses adhere closely to the original context, while factual consistency is 0.80, with minor inconsistencies. Entity accuracy is high at 0.90, reflecting accurate tagging of names, dates, and organizations. The hallucination rate is very low at 0.10, showing minimal erroneous outputs. These results highlight that the retrieval-augmented technique effectively reduces hallucinations, making the framework valuable for organizational knowledge management and decision-making.

The ablation study evaluates the impact of key components on system performance in Table 3. The Full System (with RAG, semantic embeddings, and reranking) achieved the highest performance across all metrics (Precision @10 $=$ 0.75, Recall @10 $=$ 0.85, nDCG@10 $=$ 0.92, BERTScore $=$ 0.91). Removing RAG resulted in a significant drop, reducing Precision@10 to 0.50, showing its importance for context-driven retrieval. Excluding semantic embeddings lowered performance further, with Precision@10 at 0.45 , highlighting the need for semantic understanding in document retrieval.

5. Conclusion

This article introduces a smart agent that improves retrieval capabilities and LLM power, overcoming traditional barriers in enterprise knowledge management. It effectively handles information retrieval, summarization, question answering, and recommendations using deep learning, semantic embeddings, and preprocessing. The system achieves high performance, with summarization coherence (BERTScore=0.93), retrieval recall (0.93), ranking (nDCG @ k = 0.95), and low hallucination rates (0.10) along with high entity precision (0.90). Scalability tests show the system efficiently manages large email volumes, enhancing organizational knowledge continuity and decision-making while reducing information overload.

Future research directions include domain adaptation (health, banking, law), multimodal fusion for knowledge graphs, adaptive learning for continuous improvement, the integration of explainable AI for transparency, and efficiency improvements through lean embeddings and distributed architectures for scaling deployment. These advancements will make the system more scalable, flexible, and applicable to industry needs in enterprise knowledge management.

6. Acknowledgement

7. Declarations

8. Data availability: not applicable

Conflicts of Interest: The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding Statement: This research received no external funding.

9. Author contribution

Shuo Tian: Conceptualization, methodology, data curation, writing - original draft.

Hongmei Li: Investigation, validation, writing - review and editing.

Yufei Chen: Supervision, project administration, formal analysis, writing - review and editing, corresponding author.

Ethical Approval: This study does not involve human participants or animals. Ethical approval was not required.

Consent to Participate: Not applicable.

Consent to Publication: Not applicable.

Competing Interests: The authors declare that they have no competing interests.

References

[1] A. Veglis, T. Saridou, K. Panagiotidis, C. Karypidou, and E. Kotenidis, (2022) *"Applications of Big Data in Media Organizations"* **Social Sciences *11***(9): 414. DOI: 10.3390/socsci11090414.

[2] F. Alfawaire and T. Atan, (2021) *"The Effect of Strategic Human Resource and Knowledge Management on Sustainable Competitive Advantages at Jordanian Universities: The Mediating Role of Organizational Innovation"* **Sustainability *13***(15): 8445. DOI: 10.3390/su13158445.

[3] E. Autio, R. Mudambi, and Y. Yoo, (2021) *"Digitalization and Globalization in a Turbulent World: Centrifugal and Centripetal Forces"* **Global Strategy Journal *11***(1): 3–16. DOI: 10.1002/gsj.1396.

[4] M. Aslam et al., (2022) *"Getting Smarter about Smart Cities: Improving Data Security and Privacy through Compliance"* **Sensors *22***(23): 9338. DOI: 10.3390/s22239338.

[5] G. Zhang, C. Lu, and Q. Luo, (2025) *"Application of Large Language Models in the AECO Industry"* **Buildings *15***(11): 1944. DOI: 10.3390/buildings15111944.

[6] H. Li, R. Yang, S. Xu, Y. Xiao, and H. Zhao, (2024) *"Intelligent Checking Method for Construction Schemes via Fusion of Knowledge Graph and Large Language Models"* **Buildings *14***(8): 2502. DOI: 10.3390/buildings14082502.

[7] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale, (2024) *"The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals"* **Nature Machine Intelligence *6***(4): 383–392. DOI: 10.1038/s42256-024-00820-y.

[8] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, and K. Zheng, (2024) *"When large language models meet personalization: Perspectives of challenges and opportunities"* **World Wide Web *27***(4): 42. DOI: 10.1007/s11280-024-01276-1.

[9] S. Ross, F. Martinez, S. Houde, M. Muller, and J. Weisz. "The programmer's assistant: Conversational interaction with a large language model for software development". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, 491–514. DOI: 10.1145/3581641.3584037.

[10] J. Mökander, J. Schuett, H. Kirk, and L. Floridi, (2024) *"Auditing large language models: a three-layered approach"* **AI and Ethics *4***(4): 1085–1115. DOI: 10.1007/s43681-023-00289-2.

[11] A. Kasirzadeh and I. Gabriel, (2023) *"In conversation with artificial intelligence: aligning language models with human values"* **Philosophy & Technology *36***(2): 27. DOI: 10.1007/s13347-023-00606-x.

[12] D. Anisuzzaman, J. Malins, P. Friedman, and Z. Attia, (2025) *"Fine-tuning large language models for specialized use cases"* **Mayo Clinic Proceedings: Digital Health *3***(1): 100184. DOI: 10.1016/j.mcpdig.2024.11.005.

[13] Y. Zhang and Y. Hao, (2024) *"Traditional Chinese medicine knowledge graph construction based on large language models"* **Electronics *13***(7): 1395. DOI: 10.3390/electronics13071395.

[14] D. Guo, H. Chen, R. Wu, and Y. Wang, (2023) *"AIGC challenges and opportunities related to public safety: A case study of ChatGPT"* **Journal of Safety Science and Resilience *4***(4): 329–339. DOI: 10.1016/j.jnlssr.2023.08.001.