An AI Knowledge Management System based on RAG and LLM

Erik Dethier*, Darius Hennekeuser, Daryoush Vaziri, Gunnar Stevens

Bonn-Rhein-Sieg University of Applied Sciences

{darius.hennekeuser; erik.dethier; Daryoush.vaziri; gunnar.stevens}@h-brs.de

In this paper, we demonstrate an AI-based Knowledge Management System that is intended to support new ways of knowledge interaction, especially in small organizations. Besides the efforts of leading technology providers such as Google, Microsoft, Apple, etc., our demonstration seeks to enable organizations to implement their own sovereign knowledge systems on a technological eye level. We explain the architecture of Retrieval Augmented Generation (RAG), which allows large language models (LLMs) to be extended with certain context-specific knowledge without further training. The system was co-designed in close collaboration with two companies in different consulting industries.

CCS CONCEPTS • **Human-centered computing** → human computer interaction (HCI) • **Computing methodologies** → Artificial intelligence → Natural language processing

Additional Keywords and Phrases: Knowledge Management, KMS, Artificial Intelligence, Retrieval Augmented Generation

1 INTRODUCTION

The design of Knowledge Management Systems (KMS) in organizations already has a long tradition in Human-Computer Interaction (HCI) [1–6]. However, knowledge management still faces several challenges that are now turning into exciting opportunities thanks to the evolution of AI and the growing capabilities of Large Language Models (LLMs) [7]. Knowledge management in organizational settings, such as companies, public authorities, or even private households, is about making relevant knowledge of an individual accessible and useful for themselves or others at any time in a collaborative manner. This is not just about sharing knowledge, but also ensuring that it can be understood and effectively applied by others [4, 8].

In this context, codifying knowledge into an organizational memory and ensuring its efficient dissemination and storage is still a significant challenge [7, 9]. The aim should be to conserve knowledge within the organization so that it is not going to be lost or forgotten [10], which can be a risk in, for example, commercial enterprises due to staff turnover [11] or even in private households due to needed bureaucracy-literacy of paperwork tasks [12, 13].

Emerging distributions from leading technology providers, like, for instance, Gemini from Google, CoPilot from Microsoft, ChatGPT from OpenAI, Apple Intelligence, etc., promise a revolutionary improvement in interacting with contextual knowledge, documents, emails, and much more[1]. Therefore, Artificial intelligence (AI) is currently considered to have the potential to significantly improve KMS by facilitating interaction in the capture, storage, retrieval, sharing, and

[*] Corresponding author

[1] gemini.google.com; https://io.google/2024/ (Google I/O Keynote regarding AI; access: 06/2024); copilot.microsoft.com; news.microsoft.com/reinventing-productivity/ (title: “Introducing Microsoft 365 Copilot”; access: 06/2024) openai.com/chatgpt; www.apple.com/apple-intelligence; apple.com/apple-events/ (WWDC24 regarding Apple intelligence; access: 06/2024)

application of knowledge within organizations in terms of natural language [7]. Studies show from a user perspective that AI assistants or chatbots such as ChatGPT can help to facilitate domain-specific knowledge access [14–17] and thus promote further education [18, 19]. Whereby voice-based assistants and chatbots are able to provide knowledge in real time [7]. In addition, in automatically creating and maintaining knowledge repositories, AI eases moderation tasks [20], as well as simplifies knowledge retrieval for users by automatically summarizing and categorizing relevant information and genuinely understanding user requests using natural language [7]. Furthermore, according to Pai et al. [21], AI is able to assist in capturing and disseminating tacit knowledge.

An approach that is increasingly being used to combine AI and knowledge management is Retrieval Augmented Generation (RAG), which is an innovative hybrid method that can combine LLMs with retrieved documents to fulfill context-sensitive knowledge tasks. By dynamically integrating up-to-date external knowledge without re-training an LLM, the potentials of RAG are significant [22].

In recent years, LLM innovations increasingly provided the opportunity to overcome budgetary limitations of smaller businesses and even private individuals and to enable low-cost work support as a commodity. In addition, where large enterprises have been working with complex KMS solutions for a long time and have moderated Wikis or other larger systems, leaner implementations are now also practicable, as we will show in the following. However, the handling of knowledge involves confidentiality aspects that restrict users from using solutions from (foreign/ external) LLM providers, especially in industries such as banking, which hinders the use or effective performance of knowledge work [23].

In this paper, we present an AI-based KMS that utilizes an LLM with RAG to demonstrate an innovative and usable solution for organizations. While leading technology providers such as Google, Microsoft, and Apple are striving to holistically integrate proprietary AI assistants into their operating systems and office software, we have developed a stand-alone open-source solution designed to meet the knowledge management needs of small organizations. Our prototype aims to enable organizations to create their own sovereign KMS solutions based on AI. In doing that, our system addresses the need for confidentiality of the organization's knowledge, as highlighted by Hasan & Zhou [23], by exclusively using open-source AI models in self-controlled (hosted) environments.

In designing such a system, we followed a user-centric approach and developed the system in close and participative cooperation with two small companies in the consulting sector (business management consulting and consumer counseling for household paperwork). The development involved several co-design workshops, ten interviews with employees (five from each company), and a first click-dummy testing (high-fidelity prototype) to incorporate direct user feedback. Running this approach, we ensured a usable design for this context of use.

In the following, we describe the architecture and UI of an actually realized system in detail and conclude with a discussion of (our) future work on this topic.

2 THE SYSTEM

2.1 The System's Architecture

The system has been developed as an open-source project and its components can be deployed in any (Linux) environment (Public/ Private Cloud or On-Premises) by organizations themselves or by trusted providers.

As shown in Figure 1, the architecture consists of multiple modules, which work together to build the AI-based RAG system: The knowledge base, the text-embedding model, the vector database, and the LLM. All of the modules play a distinct role in processing user prompts.

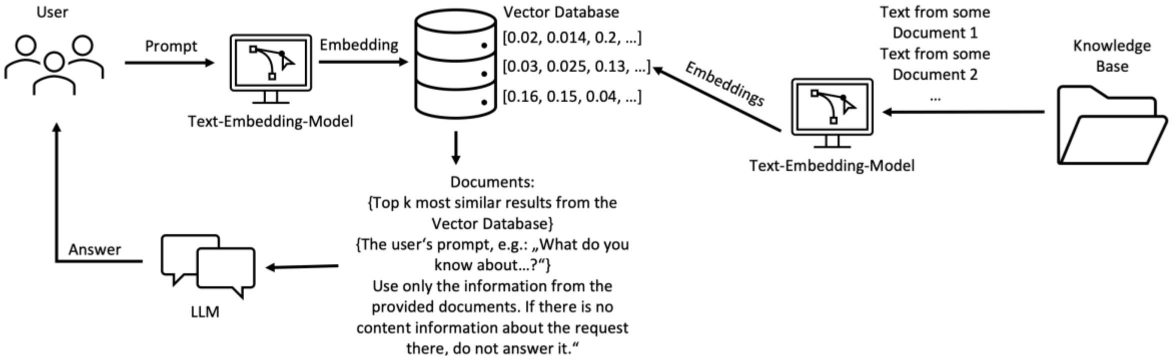

User
Prompt
Text-Embedding-Model
Embedding
Vector Database
[0.02, 0.014, 0.2, …]
[0.03, 0.025, 0.13, …]
[0.16, 0.15, 0.04, …]
Embeddings
Text-Embedding-Model
Text from some Document 1
Text from some Document 2
…
Knowledge Base
Documents:
{Top k most similar results from the Vector Database}
{The user's prompt, e.g.: „What do you know about…?"}
Use only the information from the provided documents. If there is no content information about the request there, do not answer it."
Answer
LLM

Figure 1. Overview of the System's Architecture

The textual data in the knowledge base is processed using a text-embedding-model. This model converts chunks of text into vector representations, essentially numerical representations of the text. These vector representations allow us to compute the similarity between different texts by measuring the distance between their vectors. The text-embedding model is specifically trained to generate these numerical representations, facilitating the comparison and retrieval of relevant information from the knowledge base. As a text-embedding-model, we use the jina-embeddings-v2-base-de by Jina AI, which is a German/English bilingual text-embedding-model[2].

The vector data generated by the text-embedding model, along with additional metadata such as the associated text chunk, the document name, and a unique identifier for each document, is stored in a vector database. Vector databases are specifically designed to handle large-scale and dynamic vector data, which is produced by machine learning models transforming unstructured data into feature vectors for various types of data analytics. We utilize the open-source vector database Qdrant, which supports efficient data retrieval functionalities for RAG systems[3]. This setup ensures that the system can efficiently manage and retrieve relevant information from the extensive and dynamic datasets involved.

User queries are also transformed into text embeddings. The vector database then efficiently identifies the k-nearest text chunks that are most similar to the user's query. These relevant text chunks are retrieved from the knowledge base and combined with the user's prompt. This integration provides contextual information - the relevant knowledge - from the knowledge base to the LLM, enhancing its ability to generate accurate and relevant responses.

Finally, the LLM module processes the user's prompt, now enriched with knowledge and metadata, such as document titles, pages, etc. It also includes an additional instruction for the LLM to generate responses based on the provided document texts and to include references. This ensures that the LLM delivers accurate, contextually informed answers to user queries.

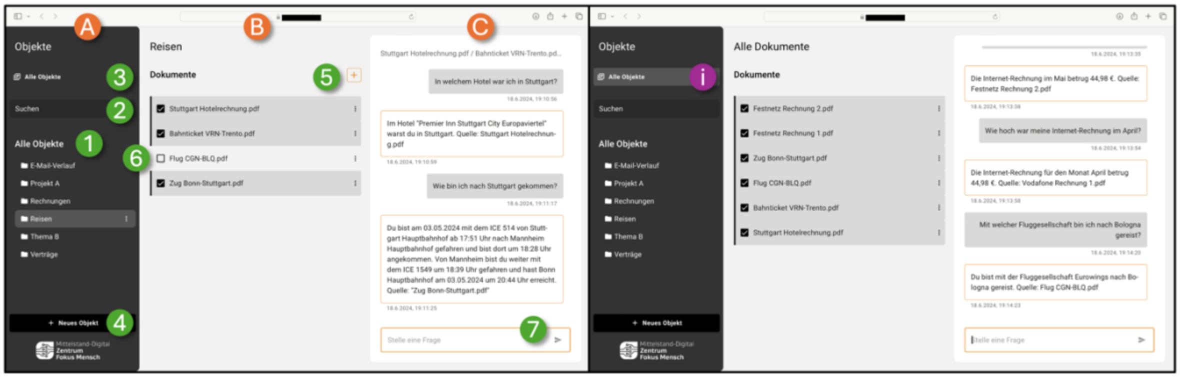
2.2 The System's UI

The UI of the system allows the user to organize knowledge within various knowledge artifacts, respectively PDF files, which the user can upload and administrate by himself (cf. Figure 2). One can organize these files thematically into a common folder-like structure that we call "Objects" (A). The files also can be selected within the Objects (B) to interact with them in a chat (C).

[2] https://huggingface.co/jinaai/jina-embeddings-v2-base-de
[3] https://qdrant.tech/rag/

As shown in Figure 2 on the left side, in the object section (A), the objects are selectable (1), searchable (2), all selectable at once (3), or new objects can be added (4). Within one object - or all objects - (B), files can be added (5) or individually selected for interaction (6). The user can select single or multiple items if he wants to use the knowledge of certain files. Interaction with the knowledge takes place within the chat based on natural language through question-and-answer. The answer is based only on text, which comes from the LLM and is enriched with high accuracy (thanks to a strong text-embedding model) exclusively by using knowledge from the knowledge base to avoid hallucinations. The full-text answer is given with the source where the knowledge comes from so that the user can check for himself.

The right side of Figure 2 shows that a user can also select all files at once (i), to search through the entire knowledge base when unsure of where the relevant knowledge is stored.



[redacted]

☑ Stuttgart Hotelrechnung.pdf
☑ Bahnticket VRN-Trento.pdf
☐ Flug CGN-BLQ.pdf
☑ Zug Bonn-Stuttgart.pdf

Stuttgart Hotelrechnung.pdf / Bahnticket VRN-Trento.pd...

In welchem Hotel war ich in Stuttgart?
18.6.2024, 19:10:56

Im Hotel "Premier Inn Stuttgart City Europaviertel" warst du in Stuttgart. Quelle: Stuttgart Hotelrechnung.pdf
18.6.2024, 19:10:59

Wie bin ich nach Stuttgart gekommen?
18.6.2024, 19:11:17

Du bist am 03.05.2024 mit dem ICE 514 von Stuttgart Hauptbahnhof ab 17:51 Uhr nach Mannheim Hauptbahnhof gefahren und bist dort um 18:28 Uhr angekommen. Von Mannheim bist du weiter mit dem ICE 1549 um 18:39 Uhr gefahren und hast Bonn Hauptbahnhof am 03.05.2024 um 20:44 Uhr erreicht. Quelle: "Zug Bonn-Stuttgart.pdf"
18.6.2024, 19:11:25

[redacted]

☑ Festnetz Rechnung 2.pdf
☑ Festnetz Rechnung 1.pdf
☑ Zug Bonn-Stuttgart.pdf
☑ Flug CGN-BLQ.pdf
☑ Bahnticket VRN-Trento.pdf
☑ Stuttgart Hotelrechnung.pdf

18.6.2024, 19:13:35

Die Internet-Rechnung im Mai betrug 44,98 €. Quelle: Festnetz Rechnung 2.pdf
18.6.2024, 19:13:38

Wie hoch war meine Internet-Rechnung im April?
18.6.2024, 19:13:54

Die Internet-Rechnung für den Monat April betrug 44,98 €. Quelle: Vodafone Rechnung 1.pdf
18.6.2024, 19:13:58

Mit welcher Fluggesellschaft bin ich nach Bologna gereist?
18.6.2024, 19:14:20

Du bist mit der Fluggesellschaft Eurowings nach Bologna gereist. Quelle: Flug CGN-BLQ.pdf
18.6.2024, 19:14:23

Figure 2: Screenshot of the System's UI (left: Interacting with selected Travel documents; right: all documents selected)

3 FUTURE WORK

In this paper, we have demonstrated the design of an AI KMS based on an LLM with RAG to prototype new ways of knowledge management and interaction in different organizations. Besides the endeavors of leading technology providers, our goal was to present a sovereign, open-source-based system that enables dedicated knowledge interaction as well as confidentiality goals in processing critical data. This enables small and medium-sized companies to implement AI themselves on an eye level with leading technology providers. Here, we have shown how to construct such a system using an RAG architecture with LLM and how a UI can look that was co-designed with users in two different companies.

However, the usability, user experience, or adoption of the system has not been extensively tested. Although the co-design approach with the two companies has provided initial valuable insights into the UI design, there is significant space for further studies to evaluate users' and organizations' situated behavior, collaboration, etc. in the field. For instance, there is a common understanding of manifold kinds of artifacts, such as paper, that are part of the knowledge work, whose integration remains an open question with respect to AI systems in general [12, 24].

Further work shall address empirical evaluations of the systems' use in various contexts respectively organizations.

ACKNOWLEDGMENTS

We thank the two companies and their employees for their support, participation, and time during the several workshops and interviews we conducted with them to develop the system. We also thank the Federal Ministry of Economics and Climate Protection of the Federal Republic of Germany for the funding.

REFERENCES

1. Evangelou, C., Karacapilidis, N.: On the interaction between humans and Knowledge Management Systems: a framework of knowledge sharing catalysts. Knowledge Management Research & Practice. 3, 253–261 (2005). https://doi.org/10.1057/palgrave.kmrp.8500076
2. Hahn, J., Subramani, M.: A framework of knowledge management systems: issues and challenges for theory and practice. (2000)
3. Hahn, J., Wang, T.: Knowledge management systems and organizational knowledge processing challenges: A field experiment. Decision Support Systems. 47, 332–342 (2009). https://doi.org/10.1016/j.dss.2009.03.001
4. Pan, S.L., Scarbrough, H.: Knowledge Management in Practice: An Exploratory Case Study. Technology Analysis & Strategic Management. 11, 359–374 (1999). https://doi.org/10.1080/095373299107401
5. Razmerita, L.: Ontology-Based User Modeling. In: Sharman, R., Kishore, R., and Ramesh, R. (eds.) Ontologies. pp. 635–664. Springer US, Boston, MA (2007)
6. Razmerita, L., Kirchner, K., Sudzina, F.: Personal knowledge management: The role of Web 2.0 tools for managing knowledge at individual and organisational levels. Online Information Review. 33, 1021–1039 (2009). https://doi.org/10.1108/14684520911010981
7. Jarrahi, M.H., Askay, D., Eshraghi, A., Smith, P.: Artificial intelligence and knowledge management: A partnership between human and AI. Business Horizons. 66, 87–99 (2023). https://doi.org/10.1016/j.bushor.2022.03.002
8. Alavi, M., Leidner, D.E.: Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly. 25, 107 (2001). https://doi.org/10.2307/3250961
9. Hansen, M.T., Nohria, N., Tierney, T.: What's your strategy for managing knowledge? In: The knowledge management yearbook 2000-2001. pp. 55–69. Routledge (2013)
10. Blackler, F., Crump, N., McDonald, S.: Organizational learning and organizational forgetting: Lessons from a high technology company. Organizational Learning and the Learning Organization: Developments in Theory and Practice, Sage, Thousand Oaks, CA. 194–216 (1999)
11. Urbancová, H., Linhartová, L.: Staff Turnover as a Possible Threat to Knowledge Loss. Journal of Competitiveness. 3, (2011)
12. Dethier, E., Kern, D.-R., Stevens, G., Boden, A.: Making Order in Household Accounting - Digital Invoices as Domestic Work Artifacts. Comput Supported Coop Work. (2024). https://doi.org/10.1007/s10606-024-09495-w
13. Emens, E.F.: Admin. The Georgtown Law Journal, 1409 (2015)
14. Ali, H., Qadir, J., Alam, T., Househ, M., Shah, Z.: ChatGPT and Large Language Models in Healthcare: Opportunities and Risks. In: 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). pp. 1–4. IEEE, Mount Pleasant, MI, USA (2023)
15. Canbek, N.G., Mutlu, M.E.: On the track of artificial intelligence: Learning with intelligent personal assistants. Journal of Human Sciences. 13, 592–601 (2016)
16. Choi, J.H., Hickman, K.E., Monahan, A.B., Schwarcz, D.: Chatgpt goes to law school. J. Legal Educ. 71, 387 (2021)
17. Sallam, M.: ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 11, 887 (2023). https://doi.org/10.3390/healthcare11060887
18. AlAfnan, M.A., Samira Dishari, Marina Jovic, Koba Lomidze: ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. JAIT. (2023). https://doi.org/10.37965/jait.2023.0184
19. Pradana, M., Elisa, H.P., Syarifuddin, S.: Discussing ChatGPT in education: A literature review and bibliometric analysis. Cogent Education. 10, 2243134 (2023). https://doi.org/10.1080/2331186X.2023.2243134
20. Jarrahi, M.H., Lutz, C., Newlands, G.: Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. Big Data & Society. 9, 205395172211428 (2022). https://doi.org/10.1177/20539517221142824
21. Pai, R.Y., Shetty, A., Shetty, A.D., Bhandary, R., Shetty, J., Nayak, S., Dinesh, T.K., D'souza, K.J.: Integrating artificial intelligence for knowledge management systems – synergy among people and technology: a systematic review of the evidence. Economic Research-Ekonomska Istraživanja. 35, 7043–7065 (2022). https://doi.org/10.1080/1331677X.2022.2058976
22. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2020)

23. Hasan, M., Zhou, S.N.: Knowledge Management in Global Organisations. IBR. 8, p165 (2015). https://doi.org/10.5539/ibr.v8n6p165
24. Briscoe, M.D.: The paperless office twenty years later: Still a myth? Sustainability: Science, Practice and Policy. 18, 837–845 (2022). https://doi.org/10.1080/15487733.2022.2146370