UDC: 004.8:004.659
Professional paper

THE ROLE OF VECTOR DATABASES IN THE ERA OF GENERATIVE AI

Merita KASA HALILI[1], Festim HALILI[2], Sarah KADRIU[3], Viola MAZLAMI[4]

[1,2]Department of Informatics, University of Tetovo, North Macedonia
[3,4]Department of Financial Mathematics, University of Tetova, North Macedonia

Abstract

The rise of modern technologies has created a need for systems that manage high-dimensional data efficiently. In this context, vector databases have emerged as a key solution, especially in applications involving generative artificial intelligence (GenAI). These databases allow fast and intelligent retrieval of relevant information by storing data as numerical vectors. However, deploying these models in real-world business contexts reveals critical limitations, notably a lack of long-term memory and reliance on a static knowledge base. To address these challenges, we explore the role of vector databases in augmenting LLMs via high-dimensional semantic vector search techniques. In particular, we posit that vector databases can act as external memory and dynamic knowledge bases for LLMs, allowing chatbots to retrieve, in real time, relevant historical interactions and domain-specific information on demand. This capability ensures that the model's responses remain contextually aware and factually grounded. The motivation for this work stems from the need for enhanced customer service, personalized interactions, and efficient knowledge management processes in enterprise AI deployments. The proposed framework integrates LLM-driven text generation with real-time vector database queries (a form of retrieval-augmented generation) to ground outputs in relevant data and maintain an extended conversational context. We will evaluate this approach through case studies in customer support and organizational knowledge management systems, assessing improvements in response accuracy, contextual coherence, and user satisfaction. By clearly articulating the synergy between vector databases and generative AI, this research aims to contribute a conceptual framework and empirical insights, as well as practical guidance for enterprise deployment at scale. We expect to demonstrate that such integration can significantly improve chatbot performance, enabling more reliable, context-aware, and tailored interactions, and thereby advancing the state of the art in AI-driven business communication.

Keywords: AI, LLMs, chatbots, vector databases, queries

1. Introduction

The development and use of vector databases has become a highly important topic in the era of artificial intelligence and machine learning. Vector databases are essential tools for storing, managing, and searching large volumes of high-dimensional data. They are widely used in machine learning, artificial intelligence, and other data-intensive applications. This section provides an overview of vector databases, focusing on their structure and common use cases.
We discuss the shortcomings of large language models (LLMs) in this work, which while delivering state-of-the-art outcomes for generative AI are limited by fixed context windows, static training materials, and the capacity to fabricate information beyond the training set. LLMs also lack long-term memory, meaning they cannot retain and recall information across sessions or access to sensitive organizational documents. All these are challenges within knowledge-intensive and real-time application where being updated and having the ability for domain-specific information matters. This study proposes a retrieval-augmented generation (RAG) model, which integrates external vector-based knowledge bases with LLMs to address limitations in context and factual accuracy.By encoding the query and the documents as a set of high-dimensional vectors and retrieving semantically aligned material, the model grounds the output of the LLM on contemporary and reliable knowledge. The process reduces the level of hallucinations, averts the knowledge cutoff problem, and enables the inclusion of private or

https://doi.org/10.62792/ut.jas.v11.i21-22.p3068

domain-specific information. We propose a generic architecture combining an LLM and a vector database that makes the model able to dynamically retrieve, cache, and leverage external information. Our solution is undergoing tests on customer service and the management of the knowledge within the company premises and proves more accurate, applicable, and adaptable without the model having the task of continuous retraining.
Contributions of this work include:
(1) seamless interoperability between a vector database and an LLM as external semantic memory;
(2) reduction of hallucination[20] by contextual grounding;
(3) assessment of the model's effectiveness on real-world, knowledge-based tasks.

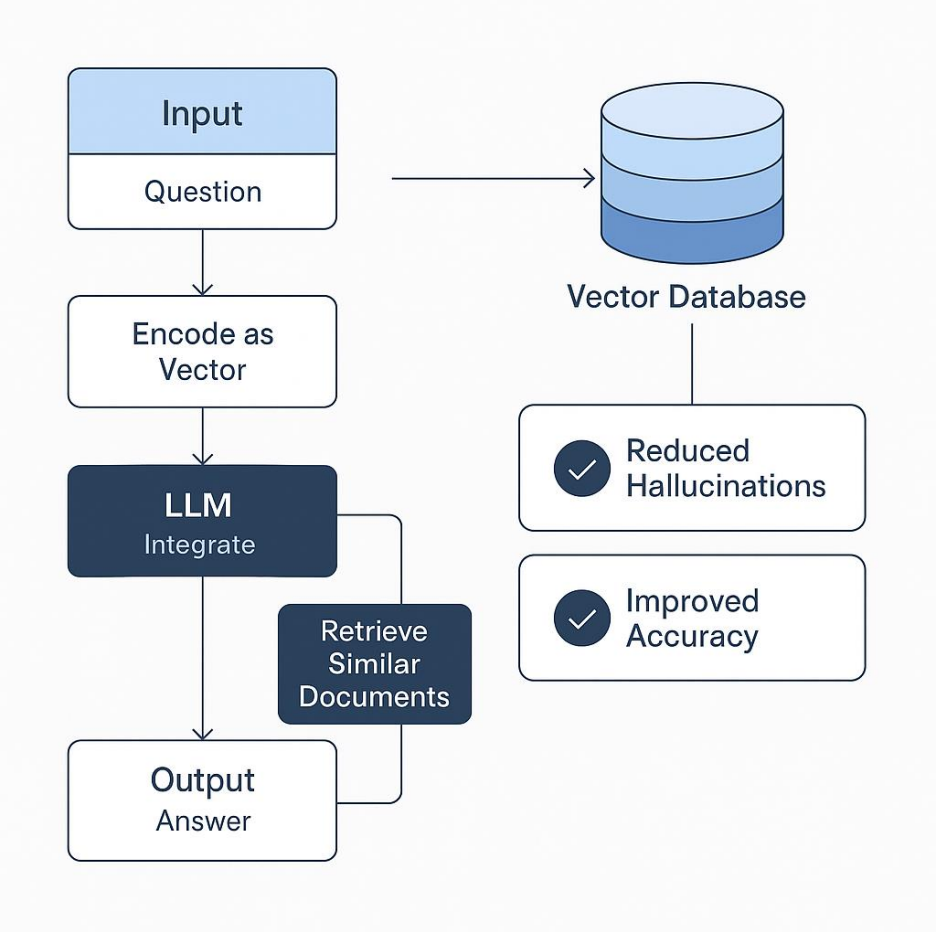


Figure 1. Architecture of LLM and Vector Database Integration (RAG)

Figure 1 illustrates the Retrieval-Augmented Generation (RAG) process, which integrates vector databases with large language models (LLMs) to enhance the accuracy and relevance of AI-generated responses. This approach addresses the inherent limitations of LLMs, such as hallucination (generating inaccurate or fabricated information) and the inability to access updated or context-specific data.
The process begins when a user inputs a query or a request for information. This input is then encoded as a high-dimensional vector to enable effective comparison with stored data. The encoded query vector is sent to a vector database, which contains pre-encoded vectors representing various documents. The system searches the database to find the most semantically similar documents, effectively retrieving content that aligns with the user's query.

[20] Hallucinations are answers or texts created by language models that look correct, but give false or made-up information in AI.

After finding the relevant documents, the retrieved information is combined with the LLM. This integration allows the model to generate a response that is not only linguistically sophisticated but also factually grounded. By leveraging external, up-to-date information, the RAG model significantly reduces the occurrence of hallucinations and enhances the reliability of the response.

The final output is an AI-generated answer that reflects both the natural language capabilities of the LLM and the data accuracy ensured by the vector database. This method proves particularly useful in knowledge-intensive applications, where being able to access current and precise information is crucial. Additionally, the process offers improved contextual grounding, making it suitable for real-world applications like customer support and organizational knowledge management.

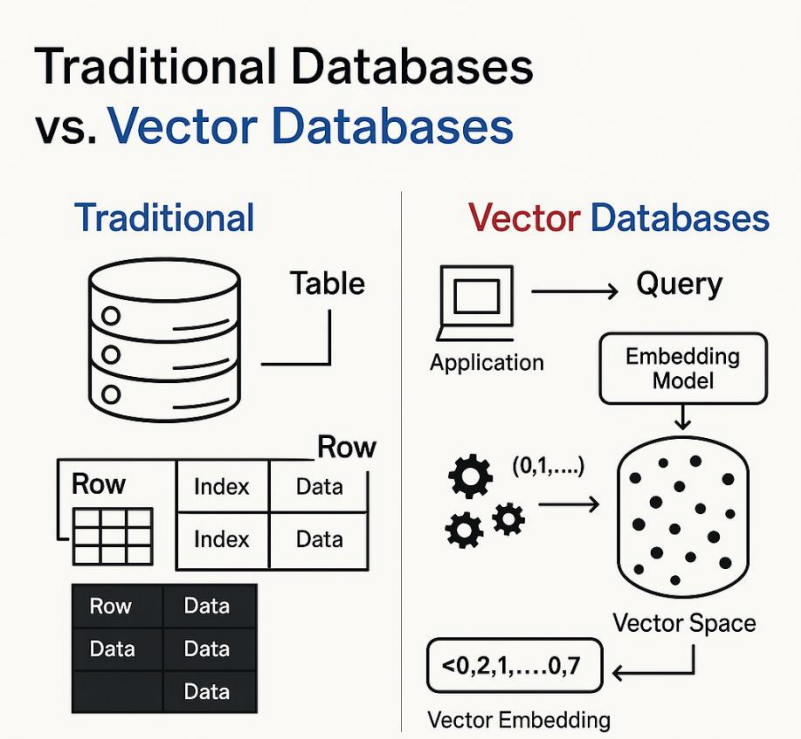


Figure 2. Comparison between Traditional Databases and Vector Databases

Most of the current systems still use classical databases to store structured data. They store the data in rows and columns in the form of tables, stock registers, or ID cards [11][15]. These databases use the SQL language to search for exact matches. It is useful where the data is the same most of the time [13][17].

But these emerging technologies demand something flexible. To address these limitations, vector databases were introduced as a flexible alternative. Unlike traditional databases that store data in structured tables, vector databases store information as high-dimensional vectors representing semantic meaning. The numerals provide the machines with the ability to understand the meaning of the data. So instead of finding exact words, the system shows results with the same meaning [12][14][16].

Prakashchand [11] mentions that the older systems make use of B-trees to search in a rapid way. Vector-based systems, however, utilize another term referred to as ANN (approximate nearest neighbor), which is more apt for more complex data [13].

An example of such a real-world application is that of the Waste Management System (WMS) by Misimi et al. [20]. It is able to measure the quantity of garbage in trash cans with the help of sensors and dispatch the corresponding trucks accordingly. This is analogous to vector search, which evaluates current conditions and selects the most optimal route.

Halili et al. [21] also applied vector search to a system which assists in locating services with the aid of intelligent, meaning-based queries. This made the system more useful as well as faster [19].
Ceri et al. [15] explain how the searches in typical systems are organized. Others, including Capella [14] and Taipalus [16], note that vector search doesn't need a rigid structure yet can support many different types of data.
Wang et al. [17], and Kale [19] indicate this is already being used in applications such as chat systems, suggestion tools, and video analysis. Daruvuri [18] further states many companies are applying the use of vector search as it is modern and powerful.

Related work

The topic of the project [1] is web-based database security with emphasis on SQL injection vulnerability. The authors refer to how individual details remain accessible even when access is not direct. They propose applying access control and encryption for greater database protection.
In paper [2] the authors explain how web services and their information can be defined using mathematical expressions in the Z-notation. The authors demonstrate how technologies such as XML, SOAP, and WSDL can be defined formally using symbols. Their aim is to make specifications of systems more precise and clear.
The author in paper [3] describes how one makes the vector databases run quicker by optimizing the algorithms that they employ. To address issues such as slow updates, high-dimensional complex data, and lack of adaptability, he advocates the employment of improved indexing, machine learning, and distributed storage.
One of the key areas of emphasis in the work [4] is the way in which vector databases handle higher-dimensional data considerably more effectively than conventional databases and how, as such, they are critical to enabling generative AI models. The work cites their use in augmenting big language models (LLMs) using Retrieval-Augmented Generation (RAG), outlines the practical applications by the big technology firms, and touches upon emerging trends such as real-time processing and integration with sophisticated AI tools such as knowledge graphs, streaming engines, and semantic retrieval systems.
In paper [5] the author talks about how vector databases can make AI models like LLMs work better. The author argues that vector databases are more efficient in handling complex, high-dimensional data compared to traditional databases. They help AI find information more quickly and accurately. He also says that big companies like Google, Microsoft, and AWS are already using this technology. He ends by talking about the future, which includes new ideas, faster systems, and making sure that AI is used safely.
The paper [6] discusses the place of vector databases as the computational backbone for embedding management in NLP, computer vision, and general AI tasks. It showcases the strengths of vector databases compared to scalar-based systems in operations such as similarity search and clustering. The paper focuses on the necessity of cloud-ready and scale-out solutions as the data scale up into the realm of billions of vectors. To ensure practicality of implementation, the author employs Python’s VectorDB for conducting accelerated similarity search.
The article [7] presents an innovative method of improved storage and query of sensor data in Green IoT networks using vector databases. By using such databases as the k-d tree and the ball tree, the technique reduces the storage capacity and speeds up the query of the data. Experimental outcomes show enhanced efficiency as well as shorter execution times compared with standard databases, and they enable real-time analysis of smart cities and healthcare.

The paper [8] overviews vector databases and embedding methods in relation to their architecture, advantages, and challenges. It describes how they transform unstructured data into vectors and discusses main indexing techniques. Major issues such as distances and dimensionality are covered as well.
The paper [9] discusses applying large language models (LLMs) on small smart devices. It presents how one should make them effectively work, securely, and responsibly. The paper also presents practical examples as well as how to construct trusted systems.
The article [10] discusses how generative artificial intelligence (GenAI) can enhance learning. It demonstrates how GenAI alters lessons and materials on the fly according to the individual needs of each student. The article discusses how GenAI might revolutionize the way we learn and teach in the future.

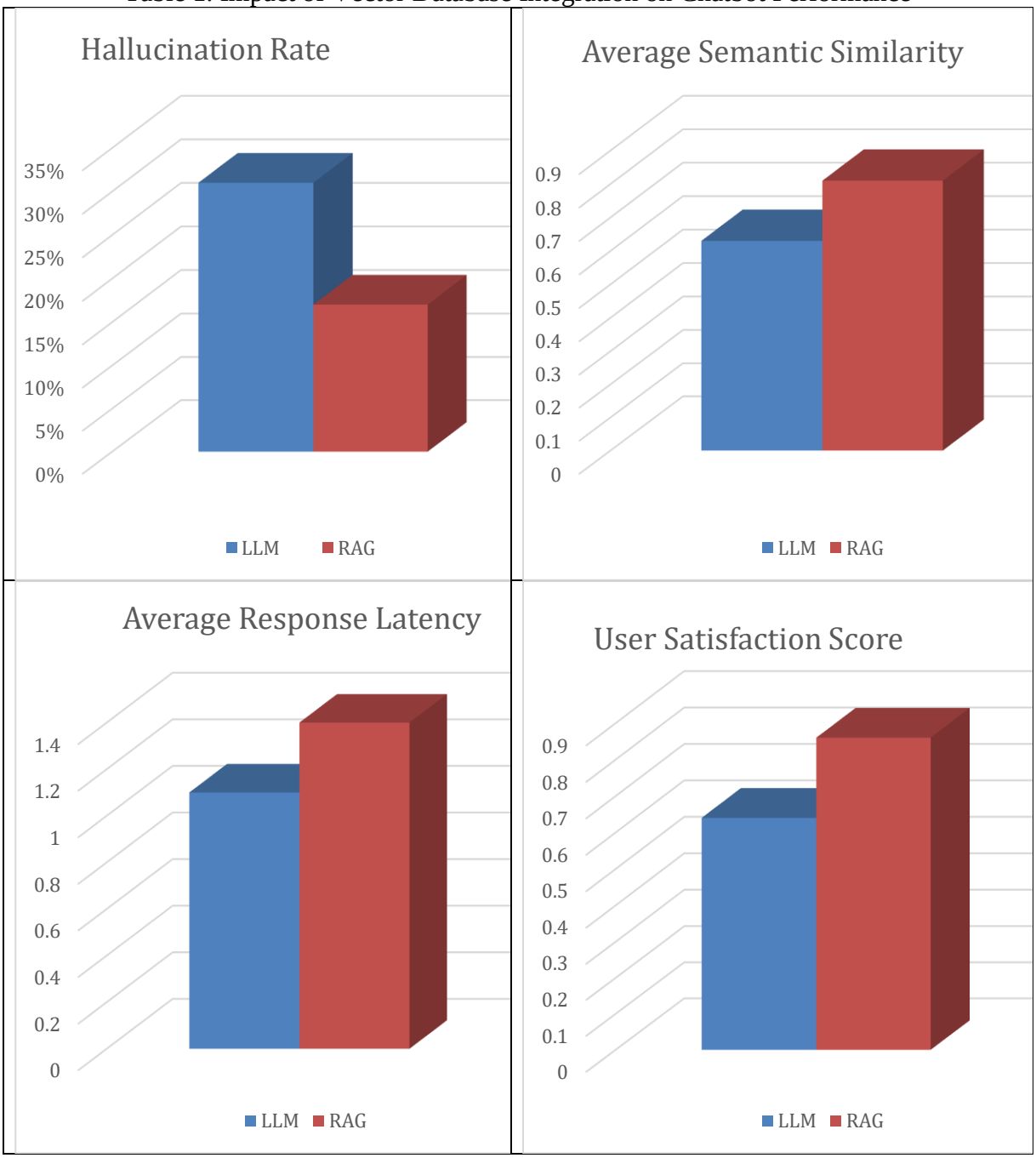
Methodology

The methodology is based on the idea that artificial intelligence can make a greater contribution and become smarter if it not only can create language but also understand how to seek out and then apply the correct information in a proper time frame.
It examines the potential that the AI system does not have to stick to the training facts alone but is also able to tap into other sources to give a truer and better answer.
Language models such as GPT-4 or LLaMA 2 are excellent at producing human-sounding sentences but are not without limitations. For instance, they lack long-term memory. As a consequence, they are not able to remember conversations or facts learned previously unless they are retrained.
They also give incorrect or outdated responses at times. It is because they are based on a static dataset and lack the capacity to update new information by themselves.
A potential solution is to develop a process through which the system can comprehend the question, look through similar or related information from various sources, and then produce a clear, accurate, and meaningful answer based on that information.
This approach reflects human-like reasoning, where information retrieval precedes decision-making.Whenever we are not sure of what we are doing, we seek facts or remember what has been previously read before responding.
This research examines cases and experiences from industry and researchers that have employed this type of strategy to enhance communication and service, e.g., customer support or the retrieval of special information.
These examples demonstrate how artificial intelligence can also grow to be more accurate and dependable if based on searchable and updateable knowledge. The aim of this methodology is to understand the notion in-depth and to demonstrate the value that it adds to the world of technology.
Rather than increasing AI systems in scale or complexity, this method suggests a simpler solution to making them intelligent by enabling them to discover what they lack when they most require it

Metric	LLM Only	LLM + Vector DB (RAG)
Hallucination Rate	31%	17%
Average Semantic Similarity	0.63	0.81

Average Response Latency	1.1 sec	1.4 sec
User Satisfaction Score	3.2/5	4.3/5
Context Retention (Multi-turn)	Low	High

Table 1. Impact of Vector Database Integration on Chatbot Performance



The comparison also demonstrates that the application of a Large Language Model with a Vector Database (RAG approach) is more effective than the application of the LLM in isolation. In specific, the hallucination rate (i.e., incorrect or fabricated responses) decreases from 31% to

17%, i.e., the model improves its accuracy. The semantic similarity score also improves from 0.63 up to 0.81, i.e., the responses are more connected to the user's query.
Although the response time is slightly longer (1.4 seconds as opposed to 1.1 seconds), the overall quality of the responses is given preference. The users' satisfaction level also improves from 3.2 to 4.3 out of 5 since the users find the system with the application of RAG to be more appropriate. The context for multi-turn dialog is also much more easily remembered with the application of RAG, and therefore, the system would be more practical for actual dialog.
Summing up, the combination of a vector database with a language model improves the accuracy, relevance, user satisfaction, and the recollection of old messages, despite the response lagging slightly.

Conclusions

The conclusion summarizes the key research questions informing the present work. According to the comparative analysis, theoretical review, and critical discussion of the data gathered in the existing literature and case studies, the following conclusions can be drawn:
Vector databases help improve the performance level of generative AI systems, especially those dealing with large language models (LLMs). Their efficient storing and retrieving high-dimensional data make them critical in supporting accurate, real-time, and context-based responses.
With the use of Retrieval-Augmented Generation (RAG) architectures, vector databases play an important part in avoiding hallucinations, the most crucial problem for contemporary AI, through the delivery of factual, up-to-date, and grounded external knowledge during the generation process.
Compared to relational database systems or classic keyword-based systems, vector-based databases allow for a richer and more flexible approach to information retrieval. It allows AI models to better interpret user intent and yield contextually accurate responses, even for vague or ambiguous queries.
Vector database adoption is on the rise in areas like customer support, healthcare, finance, and corporate knowledge management, where real-time response, accuracy, and personalization matter the most.
The combination of vector search and generative AI does not merely enhance performance from a technical point of view, but enhances user experience and trust, as it produces more credible and meaningful outputs.
At the organizational level, the adoption of vector databases in AI infrastructure can result in higher levels of innovation and efficiency, as businesses can unlock new sources of unstructured data, automate workflows for knowledge, and create smarter applications. In general, vector databases are increasingly being recognized as a strategic technological bedrock for the development of generative AI, and their value will likely increase as AI models become increasingly large-scale and complex. Emerging research and development in the near term will revolve around optimizing the storage, access speed, hybrid query systems, and privacy-sensitive retrieval algorithms for responsible and efficient use of the powerful technology.

Recommendation

This study contributes by highlighting the transformative role of vector databases in enhancing generative AI capabilities. As the conclusions reached in this work indicate, the following is recommended for the design, implementation, and responsible use of the technology in the future:

It is recommended that developers and practitioners in AI integrate vector databases into the architecture of large language models (LLMs) through methods such as Retrieval-Augmented Generation (RAG) in an effort to reduce model hallucinations, increase factual accuracy, and enable AI systems to tap into external knowledge repositories in real-time.
Data science and AI researchers should continue exploring the hybrid methods that integrate symbolic and vector-based methods for retrieval. Hybrid methods would best achieve the balance between interpretability and performance, ultimately leading to more powerful AI systems that both reason at high precision and understand the context.
Integrated into educational curricula should be an education in vector databases in machine learning, data science, and information retrieval. With the growing demand for AI skills, it is essential that the practitioners in the future gain experience in the design, usage, and optimization of the vector-based architectures. Technology firms and start-ups should invest in creating open-source, user-friendly, and enterprise-level vector database platforms that can be directly integrated into the existing AI stacks. Special attention should be given towards ensuring security, scalability, and privacy-respecting search algorithms.
Policy makers and regulators must start creating guidelines for the ethical use of vector databases, such as when they become coupled with generative AI in healthcare, education, and finance. User consent, data privacy, and transparency must be fundamental pillars that underpin such implementations.

Future work

The progress and application of vector databases is what ties the future of generative AI. As technology continues to advance, and vector databases become more powerful, they will probably be used increasingly in various applications. A number of avenues can assist further in shaping future innovation and research.
A major area of focus is the real-time integration of large language models (LLMs) and vector databases. Here, AI can fetch data from the database in real time immediately upon content generation. Thus, answers and responses made by AI shall be contextually relevant and precise.
Secondly, the present search algorithms such as HNSW (Hierarchical Navigable Small World) are performing extremely well but better ones could also be developed. Because data sizes are increasing by leaps and bounds these days, obviously, further optimization is needed. To make them faster and more efficient would go a long way in applications where quick availability of data is critical, hence speeding up the whole process.
Also, because vector databases usually hold personal and important info, security is a top priority. So, it's key to create better ways to protect and secure this data. Data privacy will greatly help in building people's trust in the tech, especially where secrecy matters.
Lastly, the use of vector databases will likely be expanded more and more across different industries. A few of them where it may help quite a lot are healthcare, education, law, and finance since these are some fields where search systems and recommendation systems greatly benefit their functionality. In these areas if the capability to quickly better accurately get the right information is through them it may enhance user experience decision making.

References

[1] Fetaji, Bekim & Halili, Festim & Fetaji, M. & Ebibi, Mirlinda. (2016). Semantic Security Analyses of Web Enabled Databases using SQL Injection. Journal of Convergence Information Technology. 11. 43-56.

[2] Halili, Festim & Rahmani, Burhan & Kasa Halili, Merita. (2012). Defining Web services and data through mathematical expressions in Z-language. 2. 51-57.

[3] Zhu, Zhongqi. (2025). Strategies for Improving Vector Database Performance through Algorithm Optimization. Scientific Journal of Technology. 7. 138-144. 10.54691/eacstn55.

[4] Joshi, Satyadhar. (2025). Introduction to Vector Databases for Generative AI: Applications, Performance,Future Projections, and Cost Considerations. IARJSET. 12. 79-93. 10.17148/IARJSET.2025.12210.

[5] Vema, Narendra Nadh. (2024). ROLE OF VECTOR DATABASES IN LARGE LANGUAGE MODELS (LLMS). INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS. 12. i436-i474.

[6] P. N. Singh, S. Talasila and S. V. Banakar, "Analyzing Embedding Models for Embedding Vectors in Vector Databases," 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, 2023, pp. 1-7, doi: 10.1109/ICTBIG59752.2023.10455990. keywords: {Analytical models;Databases;Face recognition;Vectors;Natural language processing;Data models;Behavioral sciences;Artificial Intelligence;Density estimates;Nearest Neighbor;Neural Networks;Vector databases;Vector Embeddings;Vector indices;VectorDB},

[7] R. Kumari, D. K. Sah, K. Cengiz, A. Nauman, N. Ivkovi and I. Mihaljevi, "Optimizing Resource Utilization Using Vector Databases in Green Internet of Things," 2023 IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, Malaysia, 2023, pp. 1988-1993, doi: 10.1109/GCWkshps58843.2023.10465222. keywords: {Databases;Smart cities;Green products;Medical services;Vectors;Real-time systems;Internet of Things;Green IoT;Resource Utilization;Vector Databases;Specialized Data Structures;Indexing Techniques}

[8] S. Kukreja, T. Kumar, V. Bharate, A. Purohit, A. Dasgupta and D. Guha, "Vector Databases and Vector Embeddings-Review," 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP), Yogyakarta, Indonesia, 2023, pp. 231-236, doi: 10.1109/IWAIIP58158.2023.10462847. keywords: {Surveys;Performance evaluation;Costs;Databases;Image processing;Data integrity;Data science;Artificial Intelligence;Computer Vision;Embeddings;Generative AI;Large Language Model},

[9] O. Friha, M. Amine Ferrag, B. Kantarci, B. Cakmak, A. Ozgun and N. Ghoualmi-Zine, "LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness," in IEEE Open Journal of the Communications Society, vol. 5, pp. 5799-5856, 2024, doi: 10.1109/OJCOMS.2024.3456549.
keywords: {Artificial intelligence;Computational modeling;Security;Robot sensing systems;Technological innovation;Surveys;Real-time systems;Edge intelligence (EI);generative AI;large language models (LLMs);security;privacy;trustworthiness;responsible AI},

[10] A. R. Borah, N. T N and S. Gupta, "Improved Learning Based on GenAI," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2024, pp. 1527-1532, doi: 10.1109/IDCIoT59759.2024.10467943. keywords: {Adaptation models;Technological innovation;Machine learning algorithms;Generative AI;Education;Transforms;Learning (artificial intelligence);Advanced learning;Generative AI (GenAI);Stable diffusion},

[11] Prakashchand, L. (2024). Vector Databases vs. Traditional Databases: A Deep Dive into the Evolving Data Landscape. IEEE Computer Society.

[12] Han, Y., Liu, C., & Wang, P. (2023). A Survey of Vector Database: Storage and Retrieval Technique, Challenge. arXiv.

[13] Wang, C. J., et al. (2024). Are There Fundamental Limitations in Supporting Vector Data in Traditional Databases?. PDF.

[14] Capella Solutions. 2023. Vector Databases vs. Traditional Databases: A Comparative Study.

[15] Ceri, S., Gottlob, G., & Tanca, L. (1989). What you always wanted to know about Datalog. IEEE Transactions on Knowledge and Data Engineering, 1(1), 146–166.

[16] Taipalus, Toni. (2023). Vector database management systems: Fundamental concepts, use-cases, and current challenges. 10.13140/RG.2.2.10751.79529.

[17] X. Xie, H. Liu, W. Hou and H. Huang, "A Brief Survey of Vector Databases," 2023 9th International Conference on Big Data and Information Analytics (BigDIA), Haikou, China, 2023, pp. 364-371, doi: 10.1109/BigDIA60676.2023.10429609. keywords: {Surveys;Measurement;Support vector

machines;Databases;Computational efficiency;Reliability;Indexing;high-dimensional data;nearest neighbors' search;vector databases;similarity search;similarity metrics},
[18] Da, Rajesh. (2022). Harnessing vector databases: A comprehensive analysis of their role across industries. International Journal of Science and Research Archive. 7. 703-705. 10.30574/ijsra.2022.7.2.0334.
[19] Kale, Satyanand. (2024). From Text to Recommendations: How Vector Databases are Revolutionizing Personalized Content Delivery. International Journal for Research in Applied Science and Engineering Technology. 12. 3376-3387. 10.22214/ijraset.2024.59852.
[20] Misimi, Verda & MISHKOVSKI, Igor & Halili, Festim. (2024). AN OPTIMIZATION MODEL FOR WASTE COLLECTION PATHS THAT AIMS TO CONNECT COST REDUCTION AND EMISSION MITIGATION IN ORDER TO ATTAIN SUSTAINABLE DEVELOPMENT OBJECTIVES IN NORTH MACEDONIA. Journal of Natural Sciences and Mathematics of UT-JNSM. 9. 347-353. 10.62792/ut.jnsm.v9.i17-18.p2831.
[21] Halili, Festim & Abazi, Enisa & Kasa Halili, Merita & Halimi, Halim & Bajrami, Enes. (2024). Enhancing Service-Oriented Architectures with Generative AI: A Case Study in Local Web-Based Service Discovery. Library Progress (International). 44. 730-738.