Enterprise Strategy Group™
by TechTarget

SHOWCASE

FEBRUARY 2024

Selecting a Database for Generative AI in the Enterprise

Mike Leone, Principal Analyst

Abstract: Generative AI (GenAI) and large language models (LLMs) are poised to unlock new levels of productivity, efficiency, and innovation for businesses of all sizes. Tools like embedded assistants can now handle complex, multi-step tasks, minimizing mundane work and democratizing access to knowledge. Workers are empowered to automate processes, improve decision-making, and break down information siloes.

To instill trust in generative AI across the organization, enterprises must address the pitfalls of inaccuracy, hallucination, and a lack of explainability in LLM responses. Retrieval-augmented generation (RAG) is quickly becoming the industry standard for solving these problems, but it requires a knowledge base to be effective. This report explores the requirements of a knowledge base for enterprise-ready generative AI and the limitations of vector-only approaches to propose a solution: knowledge graphs with native vector search.

Generative AI Adoption Accelerates Across Enterprises, Driving Strategic Initiatives

Among organizations that have had AI on their roadmaps for years, more than 1 in 4 organizations (26%) already have generative AI models in production, with another 66% planning to deploy in the next 12 months.[1] Research from TechTarget's Enterprise Strategy Group shows that GenAI already ranks ahead of cloud adoption on the list of strategic initiatives.[2] GenAI use cases vary widely. More than half of organizations say it will support analytics, increase employee productivity, and help improve or automate processes.[3] To date, the most common use cases include:

Market Insight

More than 1 in 4 organizations (26%) already have generative AI in production, with another 66% planning to have it deployed in the next 12 months.

- **Transforming the customer experience** with personalization through virtual assistants and intelligent call and contact centers.
- **Boosting productivity** with content creation, code generation, and summarization.
- **Improving business operations** with intelligent document processing and conversational enterprise search.

As organizations adopt generative AI, they face the challenge of realizing its short- and long-term value. Large language models, a subset of generative AI with a specialized focus on text, are painstaking and resource-intensive to build. For this reason, most companies will build their LLM applications using pre-trained foundation models. Since foundation model training focuses on generating humanlike text over fact-searching

[1] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

[2] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

[3] Ibid.

and finding, most companies will tailor foundation models to their business use case through grounding or fine-tuning.

Enterprise Challenges Driving the Need for Grounding Partners

The transition from experimental to enterprise-ready generative AI has proven challenging for many organizations. Nearly 40% cannot properly validate and evaluate results, struggle with employee hesitancy to trust GenAI recommendations, and have ethical concerns about generated content.[4]

Market Insight

Nearly 40% of organizations cite difficulty validating and evaluating results, employee hesitancy to trust recommendations, and ethical considerations and biases in generated content.

Though promising, generative AI systems are currently in the early stages of development and have yet to achieve the level of response reliability needed for mission-critical applications. Hallucinations, which occur when the LLM generates false information, prevent enterprise adoption on a broad scale. While LLMs are capable of processing large amounts of data and delivering grammatically correct responses, they cannot deliver accurate information about the business without enterprise knowledge. LLMs may also return responses that are technically correct yet lack the domain-specific knowledge that would make the response relevant and specific to the use case. For example, an LLM used to assist in legal research should be trained on existing regulations to provide the most up-to-date information.

Finally, the design of most LLMs does not include mechanisms for providing information about the data sources behind responses. Without source details, understanding how responses have been generated is impossible. Poor explainability will hinder the full adoption of LLMs and may become a dealbreaker for organizations in industries with heavier reporting requirements, such as finance or law. True explainability is difficult to achieve—it requires citing sources, explaining retrieval logic, and understanding why an LLM chooses certain pieces of information over others.

The aforementioned difficulties point to key criteria enterprises should prioritize when selecting a knowledge base for their LLM application: accuracy, relevancy, and explainability.

Trust Is the Foundation of Enterprise-grade Generative AI

Risk is a major barrier to scaling AI initiatives. Despite the hype surrounding generative AI and its transformative potential, business leaders still discourage its adoption. Executives worry about accuracy—in particular, the potential for hallucination—and believe a lack of explainability and transparency may result in poor or, worse, *unlawful* decisions. In fact, Enterprise Strategy Group research shows that just 38% of organizations measure the success and effectiveness of their AI initiatives based on the accuracy of AI models.[5] This is forcing organizations to prioritize risk management in their AI strategies to enable short-term gains and pave the way for greater ROI in the long term.

Market Insight

Just 38% of organizations measure the success and effectiveness of their AI initiatives based on the accuracy of AI models.

For organizations that neglect trustworthiness or that compromise it to market generative AI solutions quickly, the consequences can be severe. Inaccurate or irrelevant outputs may jeopardize decision-making processes, leading

[4] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

[5] Ibid.

to operational inefficiencies and financial setbacks. Furthermore, a lack of explainability compounds the problem, as it becomes nearly impossible to pinpoint and rectify the root causes.

Retrieval-augmented generation is a relatively new technique that mitigates the risks of hallucination, irrelevancy, and unexplainable results. RAG allows the LLM to tap additional data sources without retraining. Organizations can retrieve up-to-the-minute data from their own knowledge base using RAG, so that the LLM becomes a natural language interface for searching organizational knowledge.

Why Vector Falls Short in Delivering Reliable Responses

As the market for GenAI grows, the demand for a knowledge base that will support RAG has also increased. Implementing RAG requires technologies such as vector databases for rapid encoding and searching of language data to support LLM outputs. Vector databases are particularly well-suited for GenAI applications due to their unique ability to handle and process nuanced meanings in language. Unlike traditional row and column databases, vector databases encode data into multidimensional representations of data known as *vectors*. Vector distance measures the semantic similarity between words by representing them as points in space and calculating how close or far apart they are. In this space, each dimension captures an attribute of the word, such as its meaning, use, or association. For example, in a vector space the word *denim* would be positioned close to *jeans*, as denim is used to make jeans. *Cotton* might be moderately close, given denim's cotton composition, whereas *car* would be far away, since it has little relation to denim.

However, vector databases only use unstructured data with some metadata, leaving a high percentage of organizational data untapped. Most business applications of GenAI will require a well-constructed knowledge base in addition to their vector database. Aside from the problem of neglecting structured data, vector databases also fall short of delivering the deeper insights that come from inferential reasoning (also known as *knowledge-based reasoning*).

For example, consider a prompt in the asset management and financial services space about risk vulnerabilities and the asset managers likely to be most susceptible: "Which asset managers are most at risk for lithium shortage?" A vector-only LLM may surface information from reports and filings to identify general risk factors, but without knowledge of how managers have invested in different assets and businesses, the LLM could not identify the names of individual managers and the specific risks they face. This business-specific knowledge or *domain context* would have to come from a grounding partner.

The limitations of vector-only approaches to RAG become more apparent with complex prompts that involve multiple steps. Consider the question, "Which country has the highest GDP, and what is its main industry?" The vector-powered LLM may identify the United States as having the highest GDP but fail to provide correct information about its main industry. Its answer depends on vector distance rather than a search of structured data, and it may not discover the correct answer in the distance between *finance* and *technology*. Users expect responses to be accurate and relevant, so these problems can affect trust and adoption. Enterprises should ground their LLMs with a knowledge base to augment the advantages and compensate for the weaknesses of vector technology.

A Combined Approach: Knowledge Graph With Native Vector Search

Knowledge graphs are rapidly becoming the enterprise standard for LLM-based solutions that require a high threshold for accuracy. Knowledge graphs provide a structured way of representing information, which enables the LLM to draw on domain context for business-specific responses.

Global Equipment Manufacturer Using a Knowledge Graph With Vector to Improve Answers

Challenge: Support agents are taking too long to respond to customer issues. To expedite issue resolution and decrease mean time to repair (MTTR), they need help going through an extensive amount of data, including field documents, schematics, engineering summaries, documentation, and case histories.

Solution: Using a knowledge graph with vector embeddings, agents are enabled to quickly access the most relevant information and exactly match critical components to the repair, including costs, availability, expected timelines, and more.

Outcome: Agents gained efficiency in quickly delivering optimal and complete solutions that not only reduced MTTR but also instilled confidence in the company's ability to deliver on promises, improving customer experience and customer loyalty.

A knowledge graph (KG) has a graph structure: a mathematical representation where data points, called *nodes* or *vertices*, are connected by lines known as *edges*. This structure enables the KG to act as a dynamic data layer that links and organizes information from a specific domain of interest. Data is organized according to an ontology that governs the connections and classifications in the graph. Modeling data in a knowledge graph enables organizations to represent the complex, dynamic relationships between entities in the real world. A distinct advantage of knowledge graphs is their flexible schema, which allows for the inclusion of both structured and unstructured data.

Knowledge graphs facilitate the discovery of new insights through inference about the data relationships present in the graph. This inferential power complements vector databases, which cannot discern relationships in structured data. By itself, a vector database can identify the data pertinent to a query; when paired with a knowledge graph, it can leverage the relationships in a data set to inform a deeper analysis.

Consider a scenario where a user arranges travel via an LLM: "What are the airport codes for Washington, D.C.?" An LLM that depends solely on a vector database may return "DCA" or "IAD." But grounded by a knowledge graph that holds company data, the LLM can deliver a more detailed, contextualized response: "Given your coworkers' recent bookings to D.C., the most affordable economy-plus flights from your local airport are available on…" The knowledge graph contains data relationships about the employee's position, geographical location, and company guidelines on booking, all of which the LLM can draw from to generate a response. The combination of a knowledge graph, which provides inferential reasoning from data relationships, and vector search, which provides semantic similarity and understanding, delivers the most accurate, specific response.

Finally, knowledge graphs make responses more explainable by storing the relationships between data and data sources in a native data provenance layer. This enables the LLM to cite sources, provide insight into how it retrieved data, and why certain pieces of knowledge are used over others. A vector database can be used to identify the specific data source cited in its response, yet it remains unable to inspect and retrieve underlying retrieval logic as a knowledge graph can.

Conclusion

The majority of organizations consider generative AI a top strategic priority yet face challenges in transitioning from experimental stages to enterprise readiness. As risk has become the greatest barrier to unleashing the full potential of this technology, the ability to mitigate technical problems like hallucination, ambiguity, and a lack of explainability will be decisive in establishing a competitive edge in this market.

Retrieval augmented generation has rapidly become a best practice for building LLM applications tailored to specific use cases. RAG enables organizations to enhance LLM applications with another data source without modifying the underlying model, bypassing the need for resource-intensive retraining. Creating RAG applications is how most companies will scale GenAI for their business use cases quickly and at a feasible cost.

A knowledge graph with native vector search is the strongest grounding option for RAG LLM applications to date, through its powerful combination of inferential reasoning and semantic understanding. Neo4j offers the most mature knowledge graph product on the market, with unparalleled capabilities that include native vector search, multi-hop answers, cloud and ecosystem integrations, advanced algorithms, visualization, and explainability.

The selection of a database for enterprise GenAI determines the degree of sophistication that can be achieved with the LLM. The Neo4j knowledge graph with native vector search powers LLMs with advanced reasoning capabilities for a wide range of use cases, including healthcare, supply chain management, energy, customer 360, knowledge management and discovery, and more.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com
www.esg-global.com