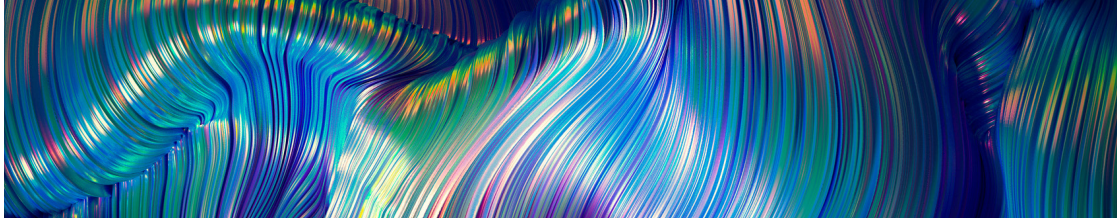
S&P Global
Market Intelligence
451 Research
Business Impact Brief



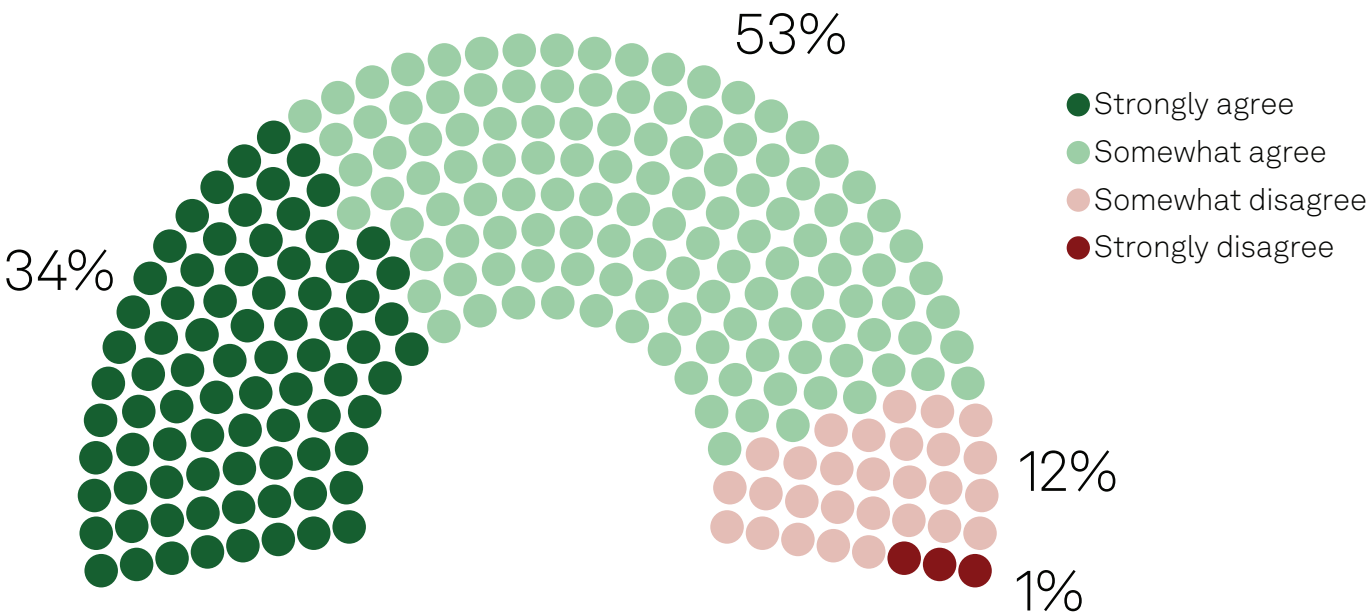
AI-generated content: Retrieval-augmented generation and why it matters

The Take

Enterprises are leaning heavily into generative AI. According to a recent 451 Research survey, 41% of respondent enterprises have adopted generative AI either in select departments or across the enterprise, and another 30% are in the process of implementing generative AI. A further 21% of respondents report informal use in their organizations, suggesting that some organizations are just dabbling with the technology.

There are tangible reasons why enterprises are excited about generative AI. The potential benefits are wide-ranging; they include enhanced productivity, improved customer service and engagement, and increased analytical insights. However, enterprises are learning that the adoption of generative AI also comes with challenges. In addition to the expense of using large language models (LLMs), inaccurate outputs, commonly referred to as hallucinations, are a concern. Hallucinations occur when a chatbot prompts an LLM on knowledge for which it has not been trained. Retrieval-augmented generation (RAG) is a way to reduce LLM hallucinations. The idea behind RAG is straightforward: It consists of augmenting an LLM with a supplemental knowledge base (in the form of a vector database) to help generate a more accurate response.

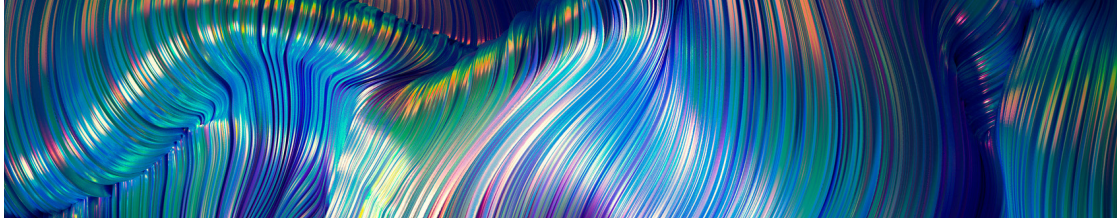
Level of agreement: A vector-supported database can be paired with an LLM to provide additional knowledge



Q. To what extent do you agree or disagree with each of the following statements? – A vector-supported database can be paired with an LLM as a way to provide additional knowledge (e.g., retrieval-augmented generation, or RAG).
Base: Organization has implemented or plans to implement vector-supported databases (n=237).
Source: 451 Research's Voice of the Enterprise: Data & Analytics, Real-Time Data Analytics and Vector-Supported Databases 2024.

S&P Global
Market Intelligence

451 Research
Business Impact Brief



Vector databases are a new breed of database that store vectors, which are numerical representations of unstructured data, such as text, images, audio and video. Vectorizing is the process of converting an enterprise's unstructured data into vectors using a specific type of LLM called an embedding model. The unstructured data is fed into the embedding model with the resulting output consisting of vector embeddings, or vectors.

The RAG environment consists of two primary components: an LLM and a vector database. Leveraging an external vector database not only provides the LLM with the necessary knowledge, but it also requires less time and expense than fine-tuning an LLM. When asked if a separate vector-supported database is a viable method to provide an LLM with additional knowledge to prevent hallucinations, 87% of enterprise leaders surveyed strongly or somewhat agreed — an overwhelming acknowledgment of their belief in RAG's benefits.

Business impact

- **Leverage domain-specific data.** While an LLM may provide general-purpose information about a variety of subjects, the data stored in a vector database can be domain-specific. Consider, for instance, an LLM-generated response on how to upsell to a particular customer. When data on the customer's purchase history, service calls and profile are retrieved from a vector store and combined with an LLM prompt, the augmented, generated response is likely to provide a richer and more accurate assessment of the next best action.
- **Reduce the use of datasets that perpetuate harmful biases and stereotypes.** One reason for biased responses is that many LLMs have been trained on public data that has little to no oversight. Biases in this data perpetuate stereotypes, which can be both offensive and harmful. RAG can help address these biases by augmenting LLM responses with enterprise data that has been curated and vetted to remove these biases, often reflecting the values of the enterprise.
- **Keep knowledge up to date.** LLMs are static entities. That is, after an LLM is trained, it requires ongoing adjustments based on multiple factors, including environmental changes and data drift. Fine-tuning is one way to update a model, but it is expensive, time-consuming and impractical to carry out on a regular basis. Updating a vector database with an enterprise's most recent data is less expensive and less time-consuming.
- **Generate responses that include citations.** It can be difficult to discern whether LLM-generated responses contain hallucinations since responses can be convincing even when they are false. RAG-generated responses can include data citations that enable users to check the source material to verify the response's accuracy. A data citation gives enterprises an added level of trust in a generated response, which ultimately provides the confidence to act based on that response.

Looking ahead

Enterprises have multiple concerns about adopting generative AI. According to 451 Research's Voice of the Enterprise: AI & Machine Learning, Use Cases 2024 survey, the top challenge in adopting generative AI is security, followed by data privacy and cost. RAG can address security and privacy, as well as trust challenges, because the data resides in a separate database, alleviating the worry that proprietary data will be included in a publicly accessible LLM. RAG can also give enterprises more confidence in the accuracy of their generated responses because the vector database that augments the LLM can be kept current with the most relevant, up-to-date data.

While RAG deployments can help quell enterprise concerns about security, privacy and cost, there are technical hurdles to consider. Not all vector databases provide the same level of functionality and performance, and pairing a database with an LLM requires technical skills. Plus, there are embedding models for generating the vectors, along with chatbots that manage the aggregated, generated responses. The future of RAG lies in a service provider's ability to abstract out the complexity of RAG such that users realize the benefits, including increased response accuracy, without the heavy lifting of managing the environment.

The promise of generative AI will only grow as more enterprises implement it into their tech stack—but its success will be limited if a business hasn't trained it using the right data. Implementing RAG will let businesses pull data into their LLM, making their outputs more transparent, more reliable, and more accurate. Simply put, it will unlock the full value of generative AI and LLMs, and give businesses the power to create consistent, dynamic, and personalized experiences tailored to their customers.