

Does AI lie ?!

HOW - AND HOW MUCH - DO VECTOR DATABASES IMPROVE AI PERFORMANCE?

Abstract

Large Language Models (LLMs) have become central in modern AI, yet they suffer from well-documented weaknesses: they produce factually wrong statements (hallucinations), their knowledge is frozen at training time, and retraining them is very expensive. This research investigates whether vector databases can address these limitations and to what extent they improve the performance of AI systems.

We reviewed recent academic papers and a dedicated benchmark comparing AI models with and without retrieval. Vector databases store unstructured data as high-dimensional embeddings and retrieve the most relevant item through Approximate Nearest Neighbor (ANN) search, allowing a model to ground its answers in external, up-to-date knowledge. This is known as Retrieval-Augmented Generation (RAG). We analysed how this approach affects the reliability, freshness, and efficiency of models outputs across different tasks.

The evidence shows that coupling a model with vector database reduces hallucinations and raises answer quality, while avoiding the cost of retraining. However, the benefit is not universal: for highly capable models, the advantage of retrieval can shrink or reverse depending on the task and the amount of context involved.

Vector databases meaningfully enhance the reliability and efficiency of AI systems, provided the retrieval strategy is matched with the use case.

Introduction

LLMs have transformed the way machines process and generate human language. Models such as GPT, Llama, and Claude can answer questions, summarise documents, and assist decision-making in fields as varied as medicine, finance, and law. Yet they share three critical weaknesses: they hallucinate, producing fluent but factually incorrect answers; their knowledge is static, frozen at training time; and updating them through retraining is expensive and often impractical.

Vector databases offer a way to overcome these limits. Rather than relying only on what the model memorised, they store data as high-dimensional vectors called embeddings and retrieve the most relevant information on demand through similarity search. Connected to an LLM, this mechanism, the RAG, lets the model ground its answers in an external, up-to-date knowledge base.

This raises the central question of our work: how, and to what extent, do vector databases improve the performance of AI systems?

Materials & Methods

This study is based on a review of recent scientific literature on vector databases and Retrieval-Augmented Generation (RAG), combined with a published benchmark measuring their effect on AI reliability.

Three types of sources were analysed: peer-reviewed academic papers describing how vector databases reduce hallucinations in Large Language Models; official documentation of vector database systems explaining how embeddings are stored and retrieved; and a public benchmark comparing model accuracy with and without retrieval.

The key figures presented in this poster come from Sng, Zhang & Mueller (2024). In their experiment, six widely used AI models were each tested on 18 questions. Every model was evaluated twice: first on its own (Before RAG), then after being given the correct source document through a vector-database retrieval step (After RAG). The proportion of wrong answers was recorded in both conditions and compared. It should be noted that the models were supplied with the correct document, so these results represent a best-case scenario for retrieval rather than an average real-world setting.

Discussions

Références

acknowledgments

Results

Giving the AI a document to read cuts its wrong answers by 72 %

