
**67% of RAG systems retrieve junk.
Because their embeddings are trash.**

Embeddings: The Hidden Backbone of RAG

**Fix your embeddings, and you fix—retrieval,
search accuracy, and hallucinations.**



Shivani Virdi

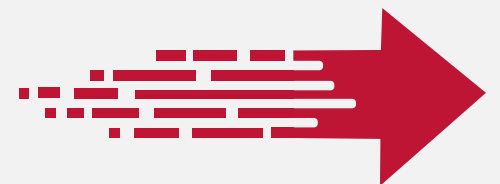


Why This Matters

- **Most devs focus on retrieval and ignore embeddings.**
- **Bad embeddings = Wrong documents = AI hallucinations.**
- **Your RAG system lives or dies by embedding quality.**



Shivani Virdi



What Are Embeddings?

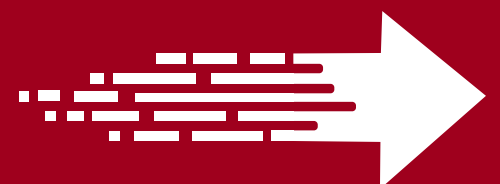
- Embeddings convert text into numbers.
- Numbers that capture meaning, not just words.

Example:

- **Search:** "best laptop for coding"
- **Without embeddings:** Returns exact matches only.
- **With embeddings:** Finds developer-friendly laptops, even if those words aren't there.



Shivani Virdi



Your RAG System Fails

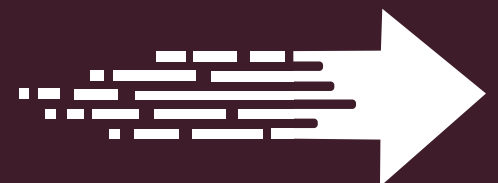
If your embeddings are weak:

- **✗ Wrong retrieval -> Hallucinations.**
- **✗ Weak context -> LLM gives bad answers.**
- **✗ Slow search -> Users get frustrated.**

Fix embeddings -> Fix your RAG.



Shivani Virdi



Hybrid Search

Most RAG systems only use dense embeddings.

📌 That's why they fail.

- **Dense embeddings** -> Understand context & meaning.
- **Sparse embeddings** -> Retrieve exact keyword matches.

👑 Hybrid search = The best of both worlds.

🚀 Precise, relevant, context-aware retrieval.



The Battle of Embedding Models

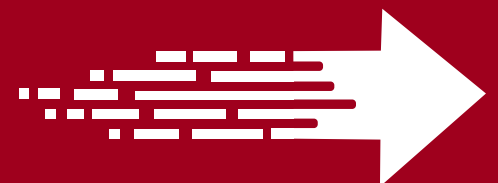
Which One Should You Use?

- **OpenAI** (text-embedding-3-small, 3-large) -> Best accuracy, but \$\$\$.
- **Cohere** (embed-multilingual-v3) -> Best for multilingual search.
- **E5 & BGE** (Open-Source) -> Free, customizable, but needs tuning.
- **Fine-Tuned Models** -> Best for domain-specific RAG, but requires expertise.

 **Breakdown in next slide -> Swipe!**



Shivani Virdi

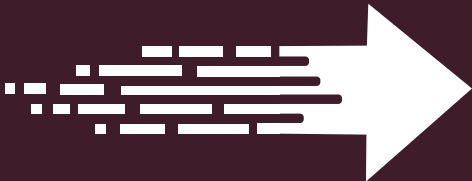


Model Comparison

Model	Strengths	Weaknesses	Best For
OpenAI (text-embedding-3-small, 3-large)	State-of-the-art accuracy, plug-and-play	Expensive	General-purpose RAG, LLM-powered search
Cohere (embed-multilingual-v3)	Multilingual support, good for global apps	Some limitations on fine-tuning	Cross-language retrieval
E5 & BGE (Open-Source)	Free & customizable, strong retrieval	Needs manual tuning	Cost-effective RAG, search-heavy apps
Fine-Tuned Models	Domain-specific precision, optimized for niche tasks	High setup cost, requires expertise	Enterprise RAG with proprietary data



Shivani Virdi




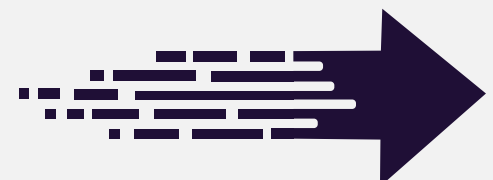
Fine-Tuning Embeddings:

Fine-tuning can **10X** your retrieval accuracy.

When should you fine-tune?

- **Legal/Medical RAG** -> Captures industry-specific terms.
- **Code retrieval** -> Understands functions/classes.
- **Customer support AI** -> Learns product-specific language.

 If generic embeddings are failing, fine-tune NOW.



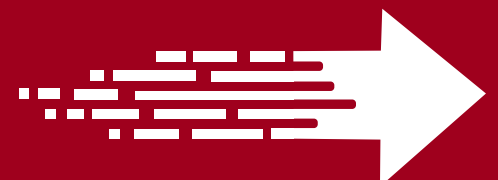
Recap

- **67% of RAG systems retrieve junk—because their embeddings are bad.**
- **Hybrid search = More precise, context-aware retrieval.**
- **Choose the right model -> OpenAI, Cohere, E5/BGE, or fine-tuned.**
- **Fine-tuning = Next-level accuracy for niche RAG systems.**

Fix embeddings -> Fix retrieval -> Fix your RAG.



Shivani Virdi





Liked This?



**SAVE
REPOST
FOLLOW**

Shivani Viridi

Which embedding
model has worked
best for you?

