[Semantic Similarity](https://www.linkedin.com/in/marie-stephen-leo?miniProfileUrn=urn%3Ali%3Afs_miniProfile%3AACoAAAQfdh4BLD4hLEVLx-k7MaGnz_cdqsDZfpI" \t "_self)

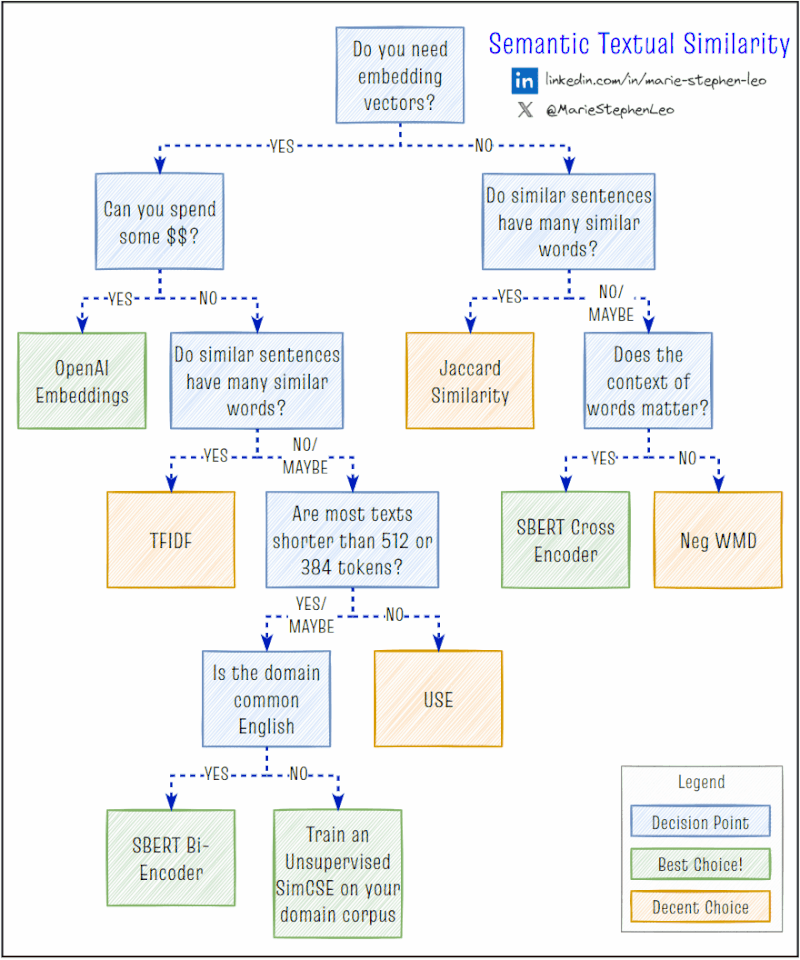
[Marie Stephen LeoMarie Stephen Leo• Follower:in• Follower:inData Director @ Sephora | Building the Cloud Data Platform (DE), LLMs, RecSys (DS/ML), and Analytics for Omni Retailing | Speaker | Writer | Instructor | Linkedin Top VoiceData Director @ Sephora | Building the Cloud Data Platform (DE), LLMs, RecSys (DS/ML), and Analytics for Omni Retailing | Speaker | Writer | Instructor | Linkedin Top Voice](https://www.linkedin.com/in/marie-stephen-leo?miniProfileUrn=urn%3Ali%3Afs_miniProfile%3AACoAAAQfdh4BLD4hLEVLx-k7MaGnz_cdqsDZfpI)

[5 Tage • 5 Tage •](https://www.linkedin.com/in/marie-stephen-leo?miniProfileUrn=urn%3Ali%3Afs_miniProfile%3AACoAAAQfdh4BLD4hLEVLx-k7MaGnz_cdqsDZfpI" \t "_self)

A 2+-year-old open-source model outperforms [OpenAI](https://www.linkedin.com/company/openai/)'s latest embedding models in Semantic Textual Similarity (STS), which is the foundation for Retrieval Augmented Generation (RAG). Still, I take back my critique from two years ago and now highly recommend OpenAI's embedding models. Here's why:  
  
In RAG systems, we transform the user's query into a numerical vector that "embeds" its context and meaning. We then compare this query "embedding" vector to the vectors of all the documents in the database to find the documents with the most similar meaning. Hence, the name semantic textual similarity.  
  
In 2022, I wrote a comprehensive Medium blog post on STS and benchmarked the various algorithms you can use. Back then, I dismissed the OpenAI embeddings as "expensive, larger in dimensions, and suffer a performance penalty compared to the best in class open-sourced models… has limited uses in the real world." Read the full article here: <https://lnkd.in/g2kgfXSu>.  
  
Last weekend, I revisited and updated the article with the latest embedding models from OpenAI. The results surprised me for two reasons!  
  
1. OpenAI's latest text-embedding-3-large and -small are worse than a 2+-year-old open-source sentence transformer model from [Hugging Face](https://www.linkedin.com/company/huggingface/) on the STSB standard dataset! Link to dataset: <https://lnkd.in/gbz2gWyA>  
  
2. The cost of generating the OpenAI embeddings has dropped by more than 99.95% in the past two years while maintaining similar performance! Latency is also reasonably quick with sub-500ms for the -small model.  
  
So now, two years since I published that post, and despite the slightly worse performance, I am happy to revise my stand to "the OpenAI embeddings are the easiest way to get started without the hassle of hosting your model inference server." I've also updated my high-level decision-making process below to reflect my changed opinion on where the OpenAI (or any managed) embedding model sits.  
  
With many more managed embedding models being accessible behind a simple API, it's easier than ever to build advanced NLP retrieval systems. However, there are some challenges with using pure semantic textual similarity that I'll discuss in my next post.  
  
Follow me for more tips on building successful ML and LLM products!  
Medium: <https://lnkd.in/g2jAJn5>  
X: <https://lnkd.in/g_JbKEkM>  
  
[#generativeai](https://www.linkedin.com/feed/hashtag/?keywords=generativeai&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224) [#llm](https://www.linkedin.com/feed/hashtag/?keywords=llm&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224) [#nlp](https://www.linkedin.com/feed/hashtag/?keywords=nlp&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224) [#artificialintelligence](https://www.linkedin.com/feed/hashtag/?keywords=artificialintelligence&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224) [#mlops](https://www.linkedin.com/feed/hashtag/?keywords=mlops&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224) [#llmops](https://www.linkedin.com/feed/hashtag/?keywords=llmops&highlightedUpdateUrns=urn%3Ali%3Aactivity%3A7166265355475124224)

Übersetzung anzeigen

Größere Bilddarstellung aktivieren,



Größere Bilddarstellung aktivieren,

* 128