<https://qdrant.tech/articles/vector-search-filtering/>

<https://qdrant.tech/documentation/guides/quantization/>

<https://qdrant.tech/documentation/guides/distributed_deployment/>

<https://qdrant.tech/articles/filtrable-hnsw/>

<https://github.com/orgs/qdrant/discussions/322>

<https://milvus.io/docs/disk_index.md>

<https://milvus.io/docs/ivf-pq.md>

<https://milvus.io/ai-quick-reference/how-does-product-quantization-pq-reduce-the-memory-footprint-of-a-vector-index-and-what-impact-does-this-compression-have-on-search-recall-and-precision>

<https://weaviate.io/blog/hybrid-search-explained>

<https://weaviate.io/developers/weaviate/concepts/search/hybrid-search>

<https://docs.llamaindex.ai/en/stable/examples/embeddings/voyageai/>

<https://www.cloudskillsboost.google/focuses/104682?parent=catalog>

<https://docs.llamaindex.ai/en/stable/examples/vector_stores/MilvusOperatorFunctionDemo/>

<https://www.googlecloudcommunity.com/gc/AI-ML/RAG-Vector-Search-Vertex-AI-Search-Grounding/m-p/867586>

<https://arxiv.org/abs/2402.15276>

<https://cloud.google.com/vertex-ai/docs/vector-search/quickstart>

<https://www.timescale.com/blog/finding-the-best-open-source-embedding-model-for-rag>

<https://big-ann-benchmarks.com/neurips21.html>

<https://arxiv.org/abs/2501.15379>

<https://docs.aws.amazon.com/bedrock/latest/userguide/titan-multiemb-models.html>

<https://www.cloudskillsboost.google/course_templates/939>

<https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/multimodalembedding>

<https://weaviate.io/developers/weaviate/concepts/vector-index>

<https://www.datastax.com/blog/best-embedding-models-information-retrieval-2025>

<https://milvus.io/docs/filtered-search.md>

<https://dev.to/mongodb/building-multimodal-ai-applications-with-mongodb-voyage-ai-and-gemini-49g3>

<https://github.com/Azure/azure-search-vector-samples/blob/main/demo-python/code/e2e-demos/azure-ai-search-cohere-embed-v3.ipynb>

<https://arxiv.org/html/2502.12799v1>

<https://github.com/GoogleCloudPlatform/generative-ai/blob/main/embeddings/intro_multimodal_embeddings.ipynb>

<https://blog.vespa.ai/billion-scale-knn-part-two/>

<https://zilliz.com/blog/json-metadata-filtering-in-milvus>

<https://github.com/AIDC-AI/Ovis>

<https://www.baseten.co/blog/the-best-open-source-embedding-models/>

<https://github.com/facebookresearch/multimodal>

<https://arxiv.org/html/2404.01292v1>

<https://openaccess.thecvf.com/content/CVPR2022/papers/Pizzi_A_Self-Supervised_Descriptor_for_Image_Copy_Detection_CVPR_2022_paper.pdf>

<https://www.researchgate.net/publication/391629733_Asymmetric_Semantic_Search_Using_Multi-Dimensional_Vector_Text_Data_Representation_Open_Semantic_Technologies_for_Intelligent_Systems_ISSN_2415-7740>

<https://innovation.ebayinc.com/stories/ebays-blazingly-fast-billion-scale-vector-similarity-engine/>

<https://milvus.io/docs/hnsw.md>

<https://milvus.io/blog/how-to-filter-efficiently-without-killing-recall.md>

<https://arxiv.org/html/2411.02537v2>

<https://www.bentoml.com/blog/a-guide-to-open-source-embedding-models>

<https://milvus.io/ai-quick-reference/how-do-i-choose-between-pinecone-weaviate-milvus-and-other-vector-databases>

<https://blog.voyageai.com/2024/11/12/voyage-multimodal-3/>

<https://www.pinecone.io/learn/series/rag/embedding-models-rundown/>

<https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-text-embeddings>

<https://aws.amazon.com/marketplace/pp/prodview-hrid2zxusacxy>

<https://qdrant.tech/articles/what-is-a-vector-database/>

<https://www.reddit.com/r/vectordatabase/comments/1kwaqx1/i_benchmarked_qdrant_vs_milvus_vs_weaviate_vs/>

<https://lancedb.com/documentation/embeddings/understanding_embeddings/>

<https://vldb.org/2024/files/phd-workshop-papers/vldb_phd_workshop_paper_id_13.pdf>

<https://m.youtube.com/watch?v=cxxEsCYt-C0>

<https://tiger-ai-lab.github.io/ABC/>

<https://www.instaclustr.com/education/open-source-ai/top-10-open-source-llms-for-2025/>

<https://blog.voyageai.com/2025/01/07/voyage-3-large/>

<https://aws.amazon.com/marketplace/pp/prodview-b4mpgdxvpa3v6>

<https://ai.google.dev/gemini-api/docs/pricing>

<https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings>

<https://sourceforge.net/software/compare/Embed-vs-voyage-code-3/>

<https://github.com/kongds/E5-V>

<https://www.researchgate.net/figure/Zero-shot-performance-comparison-of-CLIP-ALIGN-BLIP-and-CoCa-models-across_tbl2_391974842>

<https://www.geeksforgeeks.org/artificial-intelligence/understanding-blip-a-huggingface-model/>

<https://arxiv.org/html/2403.11497v1>

<https://www.reddit.com/r/googlecloud/comments/1jfk2jb/confused_about_pricing_differences_between_vertex/>

<https://liquidmetal.ai/casesAndBlogs/vector-comparison/>

<https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/multimodal-embeddings-api>

<https://www.youtube.com/watch?v=YXp6obGHy4E>

<https://redis.io/blog/searching-1-billion-vectors-with-redis-8/>

<https://dataaspirant.com/popular-vector-databases/>

<https://slashdot.org/software/comparison/Embed-vs-voyage-3-large/>

<https://github.com/milvus-io/milvus/discussions/28135>

<https://milvus.io/docs/v2.4.x/overview.md>

<https://arxiv.org/html/2403.00226v1>

<https://www.mongodb.com/blog/post/engineering/rethinking-information-retrieval-mongodb-with-voyage-ai>

<https://openai.com/index/clip/>

<https://synthesis.ai/2023/05/16/generative-ai-iv-clip-and-multimodal-retrieval/>

[https://weaviate.io/developers/weaviate/concepts/vector-index#:~:text=Hierarchical%20Navigable%20Small%20World%20(HNSW)%20is%20an%20algorithm%20that%20works,vectors%20can%20be%20resource%20intensive](https://www.google.com/search?q=https://weaviate.io/developers/weaviate/concepts/vector-index%23:~:text%3DHierarchical%2520Navigable%2520Small%2520World%2520(HNSW)%2520is%2520an%2520algorithm%2520that%2520works,vectors%2520can%2520be%2520resource%2520intensive).

<https://www.truefoundry.com/blog/best-vector-databases>

<https://roboflow.com/licensing>

<https://www.bentoml.com/blog/multimodal-ai-a-guide-to-open-source-vision-language-models>

<https://www.usenix.org/conference/fast25/presentation/tian-bing>

<https://captain-whu.github.io/SCD/SCD_files/Semantic_Change_Detection.pdf>

<https://milvus.io/ai-quick-reference/what-are-the-licensing-considerations-for-embedding-models>

<https://milvus.io/ai-quick-reference/how-do-i-choose-between-pinecone-weaviate-milvus-and-other-vector-databases>

<https://www.kdnuggets.com/the-5-best-vector-databases-you-must-try-in-2024>

<https://pureinsights.com/blog/2024/comparing-vector-search-solutions-2024/>

<https://milvus.io/id/blog/introduce-milvus-2-6-built-for-scale-designed-to-reduce-costs.md>

<https://milvus.io/docs/architecture_overview.md>

<https://milvus.io/ai-quick-reference/how-do-i-scale-my-vector-database-to-billions-of-vectors>

<https://milvus.io/ai-quick-reference/how-does-a-vector-database-handle-scaling-up-to-millions-or-billions-of-vectors-and-what-architectural-features-enable-this-scalability>

<https://docs.oracle.com/en-us/iaas/Content/generative-ai/benchmark-cohere-command-a-03-2025.htm>

<https://milvus.io/ai-quick-reference/what-vector-databases-support-multimodal-search-effectively>

<https://www.googlecloudcommunity.com/gc/Community-Blogs/Building-Enterprise-Ready-Generative-AI-Applications-with-Vertex/ba-p/723378>

<https://www.googlecloudcommunity.com/gc/AI-ML/Estimating-Vertex-AI-Vector-Search-Costs-Seeking-Cost-Effective/m-p/796972>

<https://cohere.com/blog/multimodal-embed-3>

<https://zilliz.com/learn/top-10-best-multimodal-ai-models-you-should-know>

<https://arxiv.org/abs/2505.20764>

<https://cohere.com/blog/int8-binary-embeddings>

<https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/how-to/image-retrieval>

<https://roboflow.com/model-feature/image-embedding>

<https://www.reddit.com/r/LangChain/comments/1blfg7i/what_is_the_current_best_embedding_model_for/>

<https://arxiv.org/abs/2501.13920>

<https://github.com/shanyu-sys/Awesome-multi-modal-embeddings>

<https://learn.microsoft.com/en-us/azure/ai-studio/how-to/deploy-models-cohere-embed>

<https://docs.anthropic.com/en/docs/build-with-claude/embeddings>

<https://news.ycombinator.com/item?id=43694546>

<https://www.byteplus.com/en/topic/516148>

<https://www.reddit.com/r/OpenAI/comments/yv1hir/clip_vs_blip/>

<https://www.timescale.com/blog/vector-database-basics-hnsw>

<https://milvus.io/ai-quick-reference/how-does-indexing-work-in-a-vector-db-ivf-hnsw-pq-etc>

<https://en.wikipedia.org/wiki/Hierarchical_navigable_small_world>

<https://cloud.google.com/vertex-ai/docs/vector-search/overview>

<https://jina.ai/embeddings/>

<https://docs.spring.io/spring-ai/reference/api/embeddings/vertexai-embeddings-multimodal.html>

<https://www.edenai.co/post/top-free-image-embeddings-tools-apis-and-open-source-models>

<https://www.googlecloudcommunity.com/gc/AI-ML/Confused-about-pricing-differences-between-Vertex-AI-and-Google/m-p/887859>

<https://stackoverflow.com/questions/79559493/how-to-set-up-google-vertex-ai-vector-search-with-private-endpoint>

<https://www.cloudraft.io/blog/top-5-vector-databases>

<https://milvus.io/>

<https://zilliz.com/comparison/weaviate-vs-qdrant>

<https://neuroflash.com/blog/google-vertex-ai-pricing-pros-cons/>

<https://github.com/jina-ai/jina-clip>

<https://huggingface.co/JinaAI>

<https://github.com/mlfoundations/open_clip>

<https://laion.ai/blog/laion-5b/>

<https://arxiv.org/abs/1908.04725> *(HNSW original paper)*

<https://arxiv.org/abs/1408.2927> *(Product Quantization for NN Search)*

<https://arxiv.org/abs/1702.08734> *(DiskANN for billion-scale on SSD)*

<https://vector-database-benchmark.github.io>/ *(independent ANN benchmarks)*

<https://milvus.io/docs/benchmarks.md>

<https://weaviate.io/developers/weaviate/scaling/replication-and-sharding>

<https://qdrant.tech/documentation/internals/performance-benchmarks/>

<https://pinecone.io/learn/vector-database-design-multi-tenancy/>

<https://www.pinecone.io/learn/vector-database-indexes/>

<https://zilliz.com/blog/vector-index-selection-guide>

<https://romaindelaby.github.io/posts/vector-databases-and-metadata-filtering/>

<https://medium.com/qdrant/building-billion-scale-semantic-search-with-qdrant-and-laion-clip-521a77b3611e>

<https://weaviate.io/blog/vector-search-at-scale>

<https://towardsdatascience.com/choosing-the-right-vector-database-a-practitioners-guide-7a88d2e597c9>

<https://www.milvus.io/blog/milvus-vs-qdrant-vs-weaviate>

<https://github.com/facebookresearch/faiss>