Optimize and Evaluate Embeddings to Improve Performance

Cătălin Tudose

Java Champion, PhD in Computer Science, Java and Web Technologies Expert

www.catalintudose.com | www.linkedin.com/in/catalin-tudose-847667a1

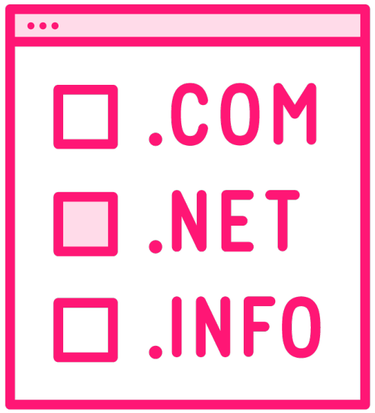
Why Evaluation of Embeddings Matters

Quality assurance

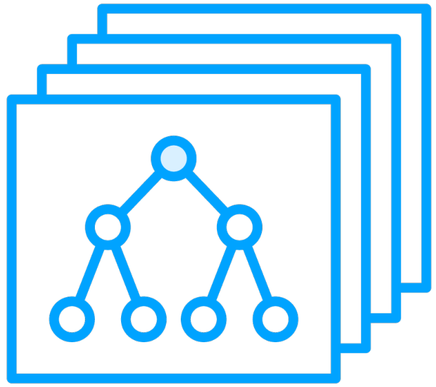
Performance measurement

Optimization

Why Not All Embeddings Are Equally Effective



Domain mismatch



Task dependence

Representation limits

Intrinsic Evaluation

Word similarity tests

Analogy tasks

Clustering and visualization

Extrinsic Evaluation

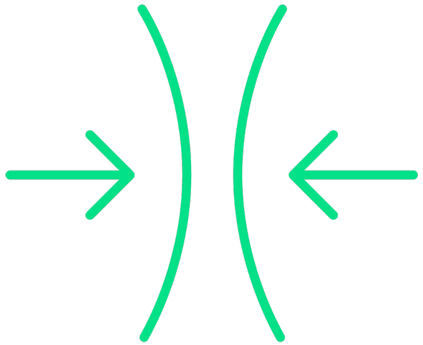
Search relevance

Classification accuracy

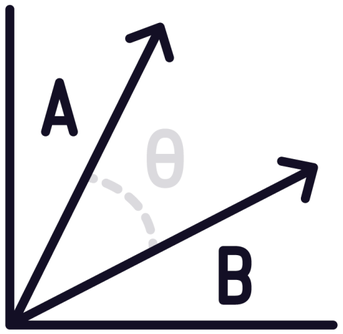
Question answering performance

Task-driven perspective

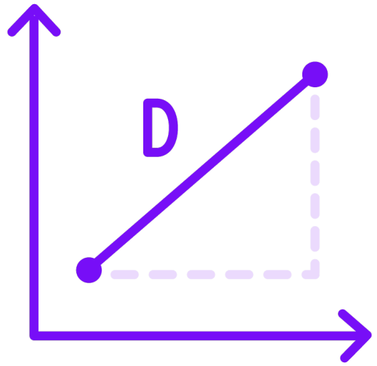
Similarity Metrics



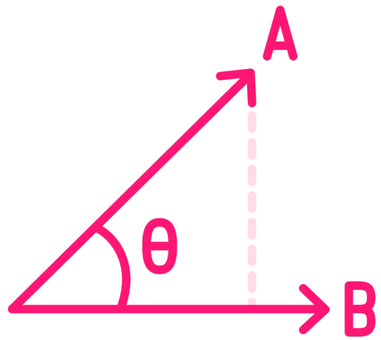
Closeness in vector space



Cosine similarity

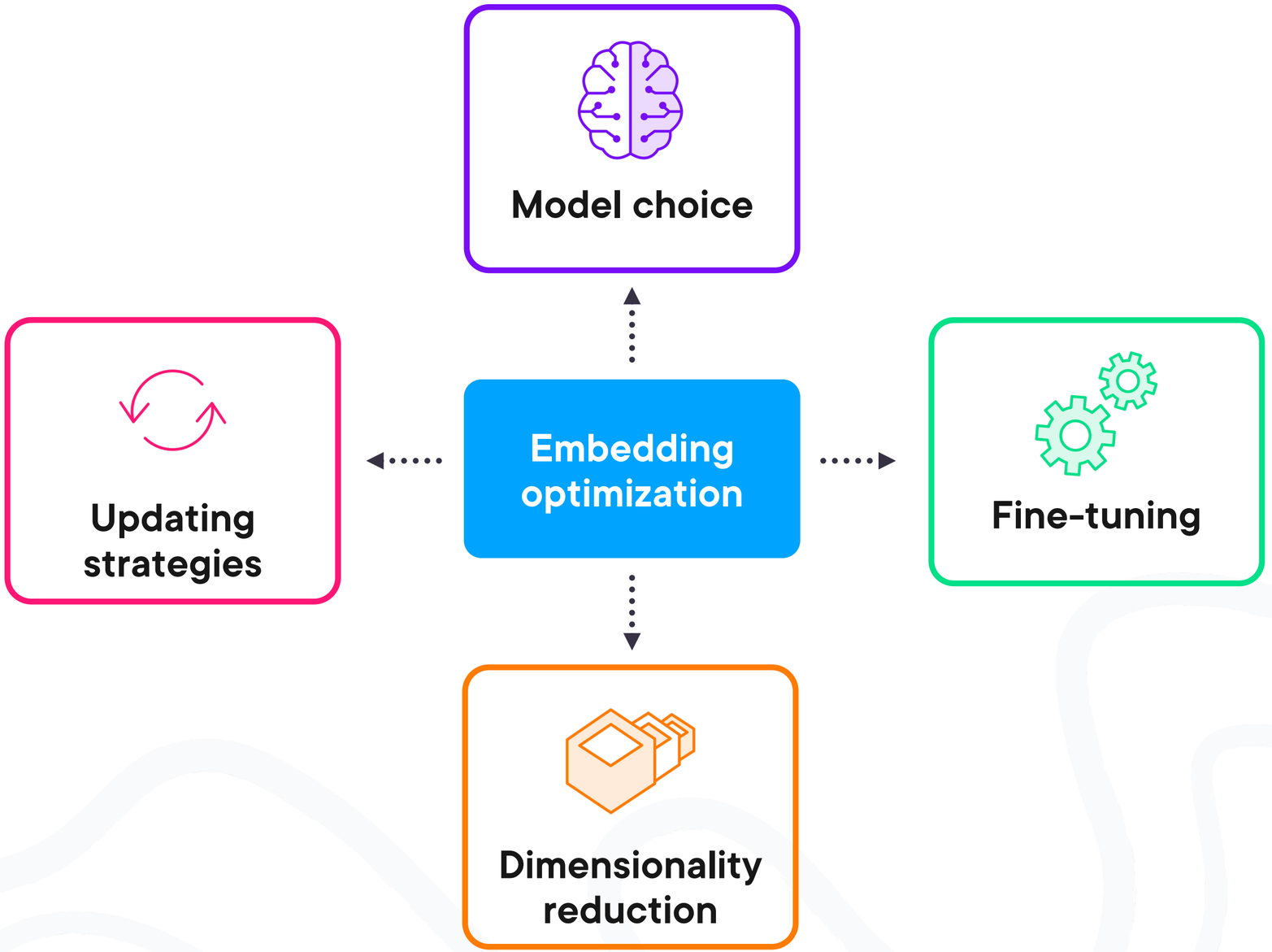


Euclidean distance



Dot product

Optimization Levers



Choosing the Right Embedding Model

Varying accuracy and efficiency

Larger models, better semantic coverage

Higher computational and storage costs

Smaller models, faster but less precise

Fine-Tuning

Adapting to a specific domain improves relevance

The vector space better reflects specialized terminology

Increases retrieval accuracy for domain-specific queries

Dimensionality Reduction

High-dimensional vectors consume storage

Compress embeddings

Speed up vector search

Lower memory requirements

Minimal accuracy loss

Updating Strategies

Incremental updates

Periodic refreshes

Hybrid approaches

Keep retrieval pipelines aligned with the latest information

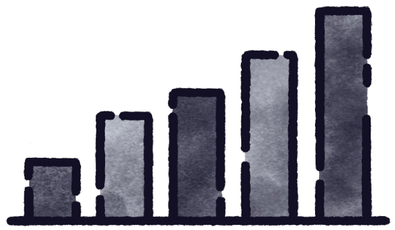
Bias Concerns in Embedding Generation

Source of Bias

Embeddings learned from large text corpora

Stereotypes, offensive associations, cultural imbalances

Impact of Bias



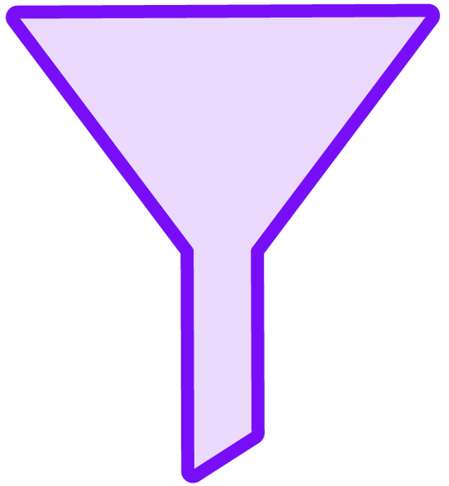
Search and ranking

Downstream LLM responses

Fairness & Trust

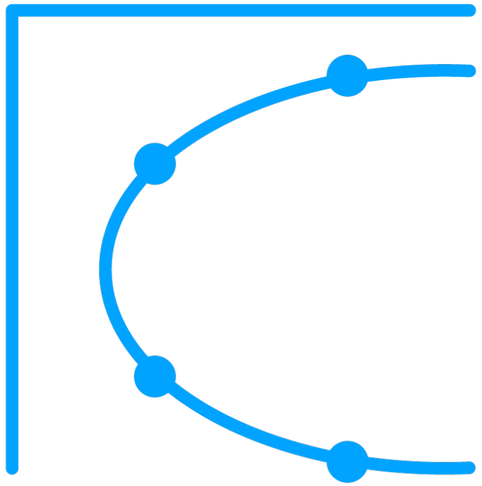
Data Curation

Clean, balanced, diverse corpora

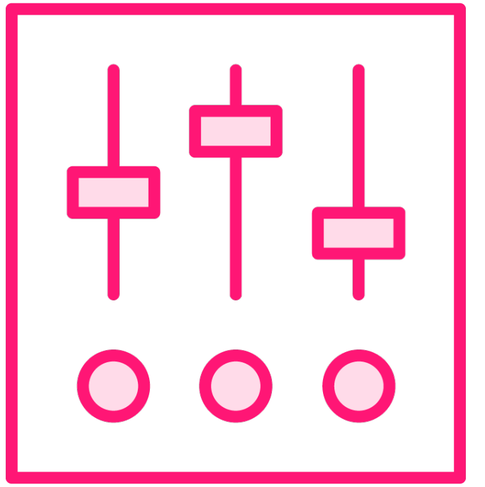


Filter out toxic, biased, or underrepresented language

Model Choice & Fine-tuning



Models trained with fairness considerations



Fine-tune embeddings on domain-specific data

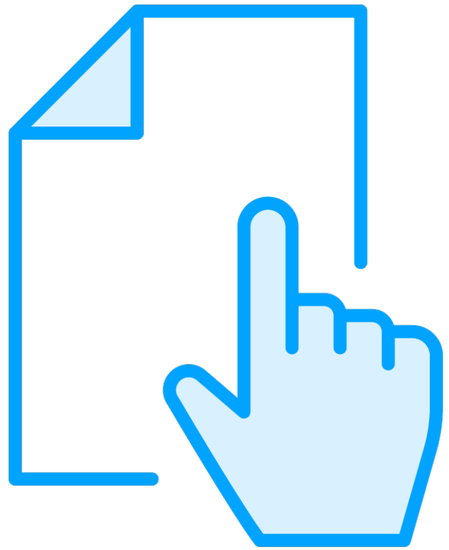
Concerning Monitoring & Evaluation

Use bias evaluation benchmarks

Incorporate human-in-the-loop review

Transparency & Documentation

Document the limitations and known biases



Provide usage guidelines for developers

Thank you for your time!

www.catalintudose.com | www.linkedin.com/in/catalin-tudose-847667a1