FineTune EmbeddingGemma and run on CPU

THỬ NGHIỆM

Google AI Developers

EMBEDDING
GEMMA

NGON, BỔ, RẺ?

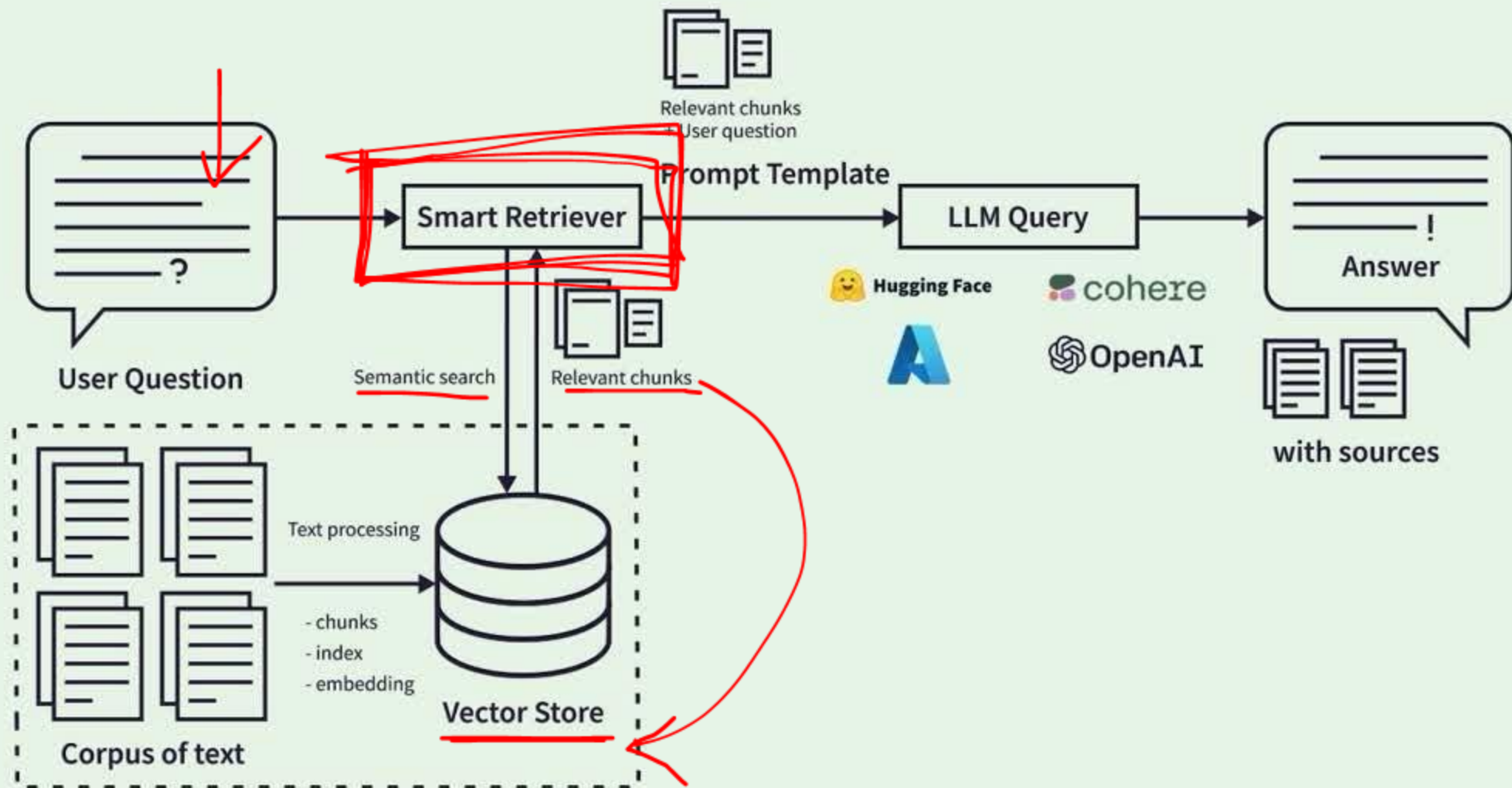so găng với các model embedding dangvantuan, aiteamvn...

miai.vn

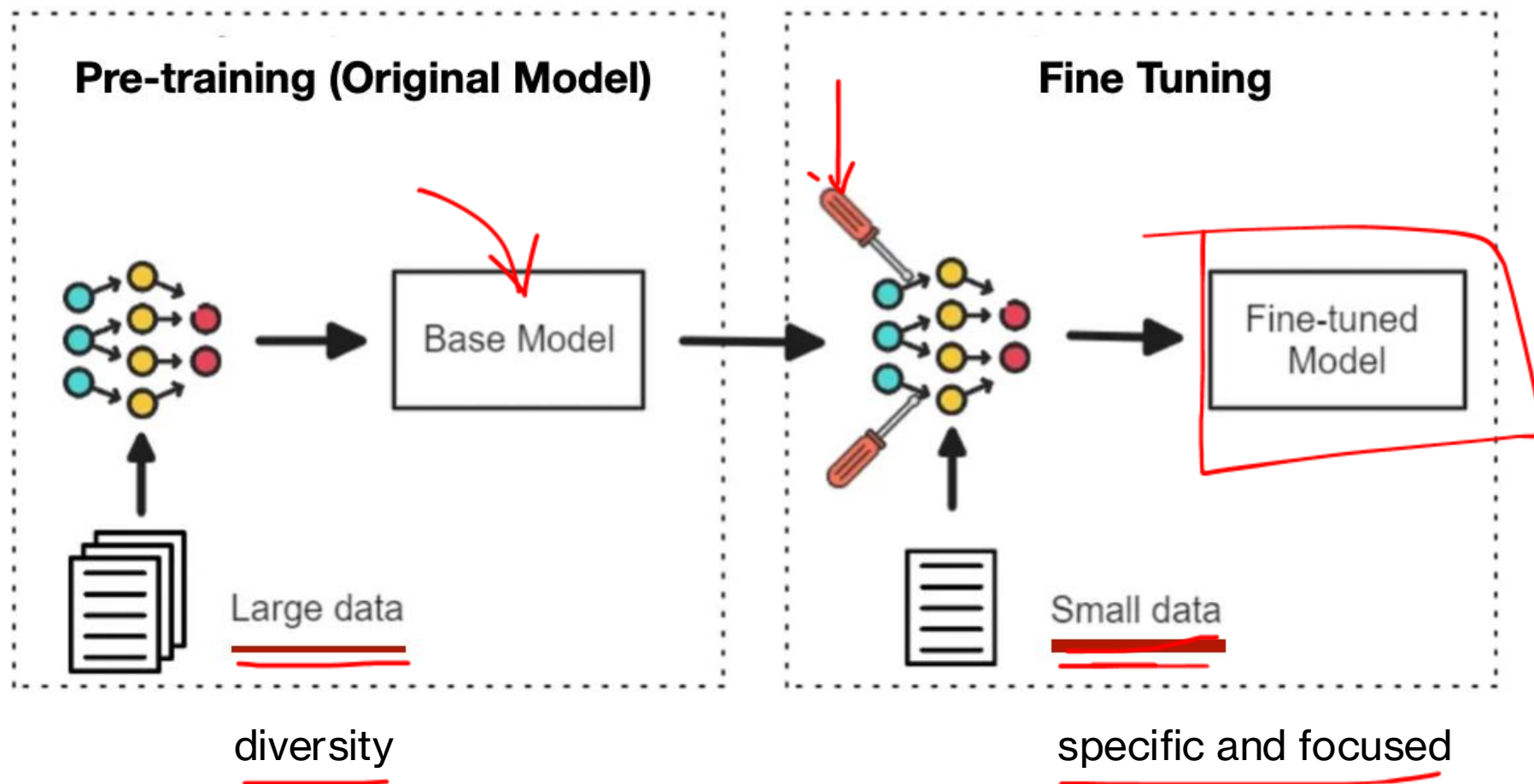https://www.youtube.com/watch?v=vmDSCNWMa9c

Why finetune embedding model?

Retrieval Augmented Generation (RAG)

Why finetune?

Large Language Model

Pre-training (Original Model)	Fine Tuning
Base Model	Fine-tuned Model
Large data	Small data

diversity   specific and focused

Before & After

LinkedIn Jobs

	Before	After

Query: "Business Analyst"

Near: "A Business Analyst bridges the gap between IT and the business, ensuring that technological solutions align with business needs by analyzing data and documenting requirements"

Before: 0.5

After: 0.7

Far: "Construction Worker's role is to physically build and assemble structures, focusing on manual labor, machinery operation, and on-site problem-solving to complete projects"
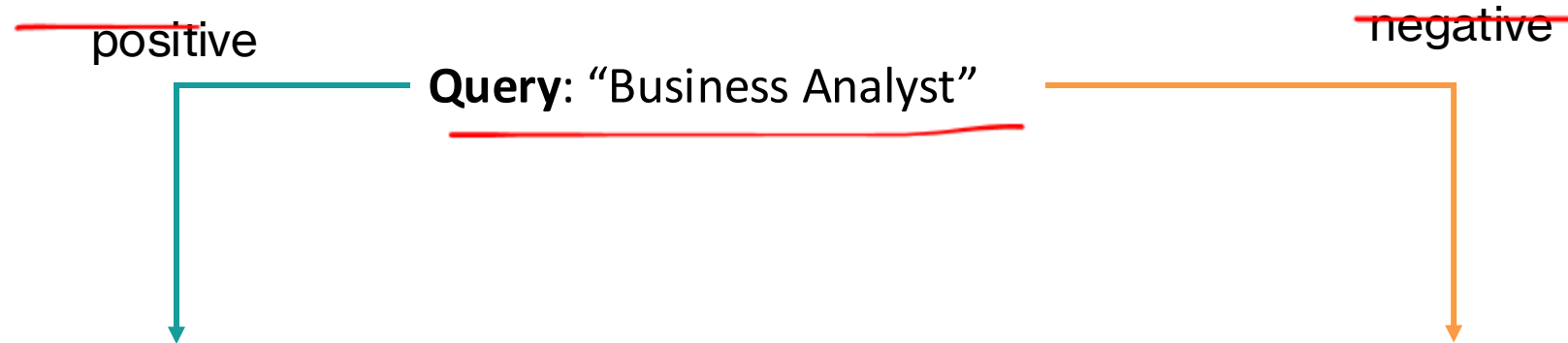
Before: 0.3

After: 0.1

Dataset

positive   negative
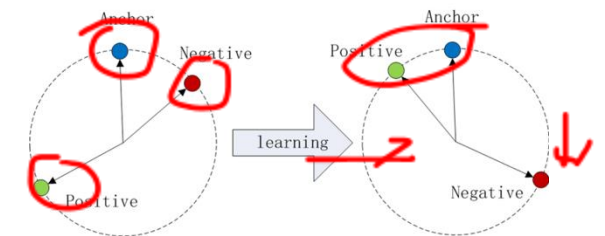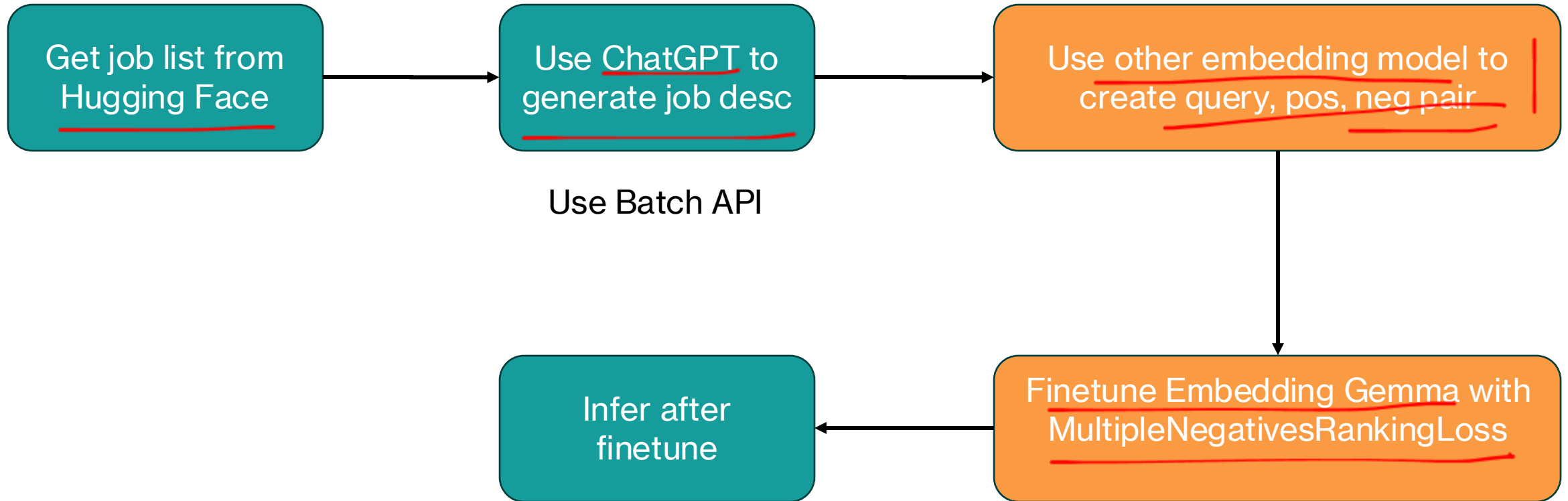
Query: "Business Analyst"

Near: "A Business Analyst bridges the gap between IT and the business, ensuring that technological solutions align with business needs by analyzing data and documenting requirements"

Far: "Construction Worker's role is to physically build and assemble structures, focusing on manual labor, machinery operation, and on-site problem-solving to complete projects"

Flow

Get job list from Hugging Face

Use ChatGPT to generate job desc

Use other embedding model to create query, pos, neg pair

Use Batch API

Infer after finetune

Finetune Embedding Gemma with MultipleNegativesRankingLoss

Hands-on