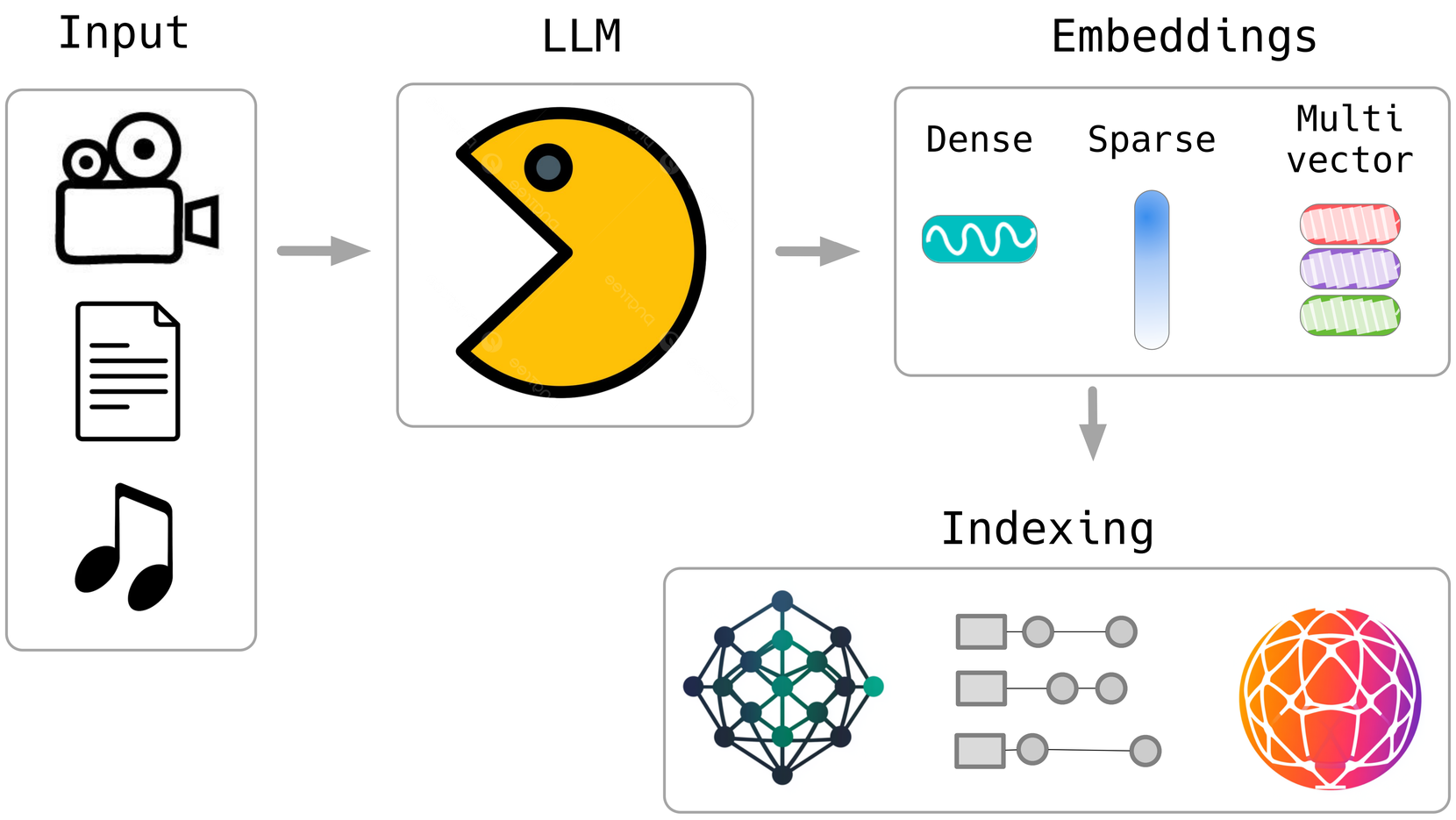
Approximate Nearest Neighbors Search over Neural Embeddings



Topics

LLM

Large Language Models **encode** different kind of inputs into vectors. **Complex** relationship are modeled by means of **simple** similarity metrics.

Model Compression

Many tasks, including information retrieval and recommendation systems, are **solved** by **k-Nearest Neighbors (k-NN)** search.

KNN Search

Exact k-NN -> unfeasible. **Approximate** solutions, including **graphs**, **inverted indexes**, **quantization**, are waiting for you to improve them!

Work with us!

Model Compression

Exploit **Model Compression** techniques reduce the large computational burden of LLM without affecting performance.

Encoders Machinery

Invent new **training algorithms** and/or embeddings format to improve the effectivness of LLM-based encoders.

Embeddings Quantization

Reduce the memory footprint and the retrieval time with quantization:

1. Lossless, algorithmic.
2. Lossy, machine learning powered.

Index Design

Design data structures for efficient and effective nearest neighbors search.

KANNOLO library

Contribute to the fastest available ANN library, written in Rust.

HPC lab

UNIVERSITÀ DI PISA

TusKANNy

silvio.martinico@phd.unipi.it
francomaria.nardini@isti.cnr.it
cosimo.rulli@isti.cnr.it
rossano.venturini@unipi.it