Dlib-Transformer-extensions

Transformers and language models for Dlib, from the import of a model to its delivery

The library on which everything we publish rests. This temporary fork of Dlib, the open-source C++ machine learning library, is the place where each mechanism is designed, demonstrated by a runnable example, measured and stabilised before it is proposed to Dlib itself.

License **Boost Software License 1.0** C++14 **CUDA, OpenVINO**

24
example programs, one per mechanism

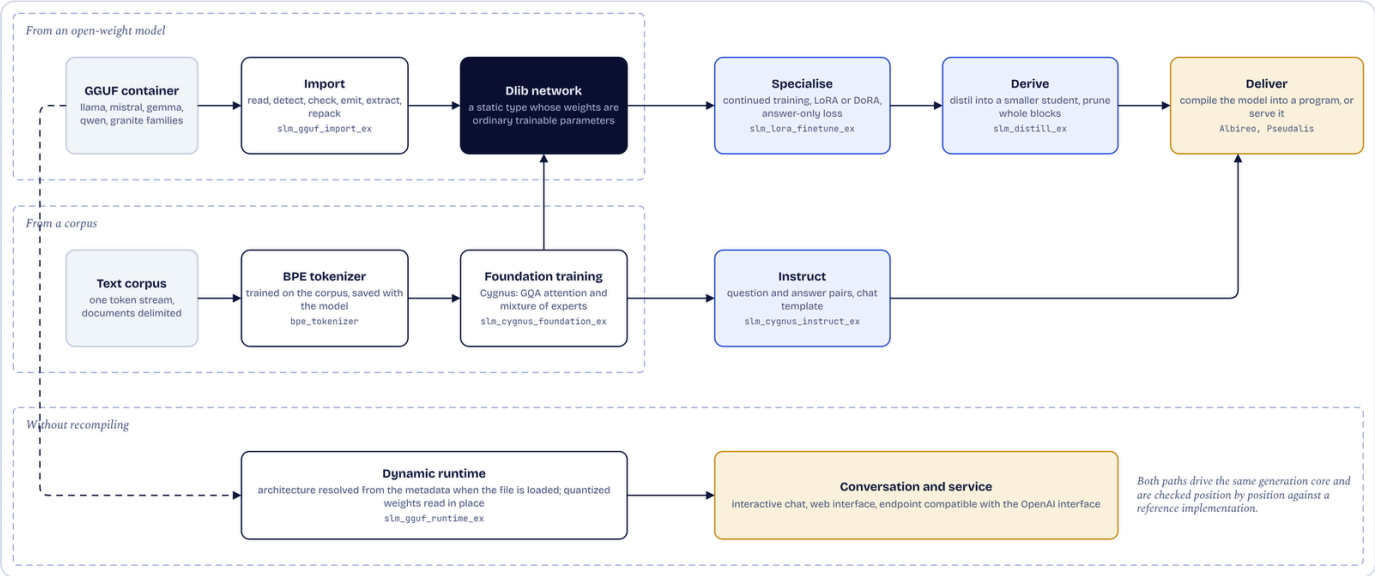
14
contributions already merged into Dlib

12
model architectures read from GGUF files

0
Python dependency in the chain

One library for the whole life of a model

An open-weight model enters through its GGUF file and becomes an ordinary Dlib network; a model of our own starts from a corpus. Both then follow the same path: specialise, derive, deliver. A second engine runs any supported model without recompiling.



Scope of the library

Vision and multimodality
Vision Transformers as image backbones or as the vision tower of a model that reads text and images, trainable end to end.

Personal data protection
Rules and small learned recognizers find personal data in five languages; format-preserving encryption replaces it reversibly.

NPU and integrated GPU
Each tensor operation runs on the device measured to be fastest: the host, the Intel NPU or the integrated GPU through OpenVINO.

Quantized weights on the card
CUDA kernels multiply activations with weights kept in their GGUF formats, without expanding them.

Retrieval
Embedding models index a folder of documents and answer a question with the passage that matches it.

Extended GPU memory
V²RAM streams blocks between the card, host memory and a disk store, so that a model larger than the card trains on it.

Processor-specific kernels
Code paths for AVX2, AVX-VNNI and AVX-512 VNNI are chosen at run time from what the processor offers.

Prediction as compression
A trained network and an arithmetic coder form a lossless compressor that any processor reproduces bit for bit.

More on
cydral-technology.eu/en/research/library/

Source code
github.com/Cydral/Dlib-Transformer-extensions

Contact
cydral-technology.eu/en/contact/

Free and open source under the Boost Software License 1.0. © 2026 Cydral Technology, France.