Cydral Technology

Dlib-Transformer-extensions

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

Cydral Technology, November 2026

The library at a glance

A temporary fork of Dlib, the open-source C++ machine learning library. Each mechanism is designed, demonstrated by a runnable example, measured and stabilised here, then proposed to Dlib itself.

C++14	Boost Software License 1.0
No Python in the chain	CUDA and OpenVINO

The path of each mechanism

Design → **Example** → **Measure** → **Stabilise**

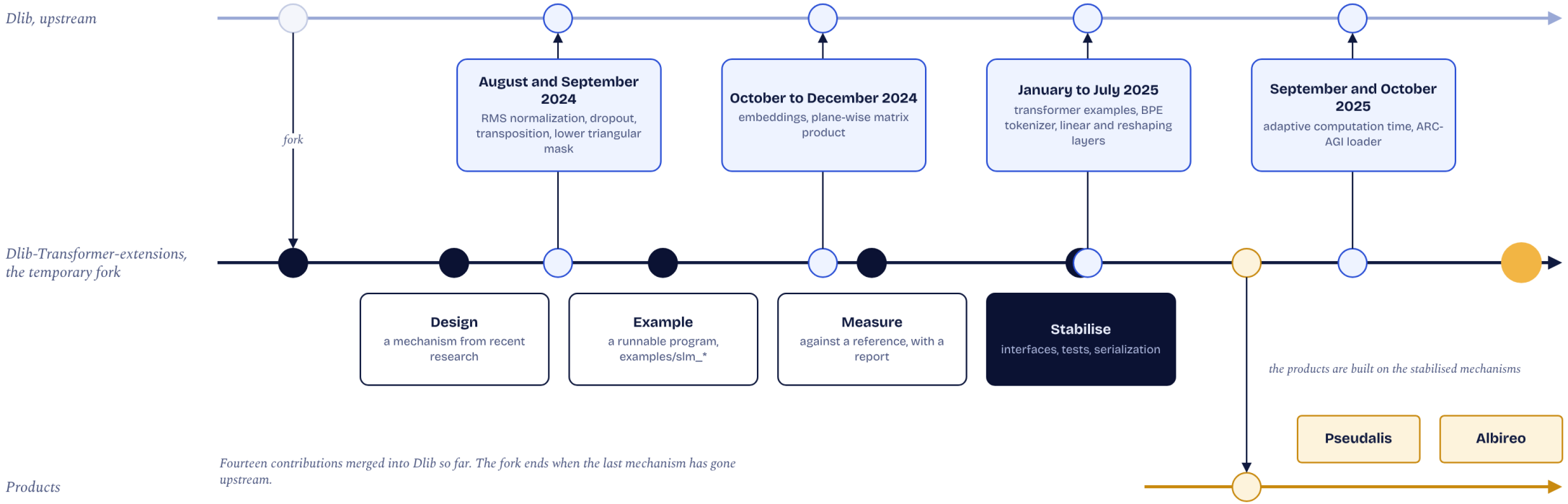
24 example programs, one per mechanism

14 contributions already merged into Dlib

12 model architectures read from GGUF files

0 Python dependency in the chain

A temporary fork that feeds Dlib and the products



Fourteen contributions merged into Dlib so far. The fork ends when the last mechanism has gone upstream.

Why a fork

Language models move faster than a reference library can accept additions.

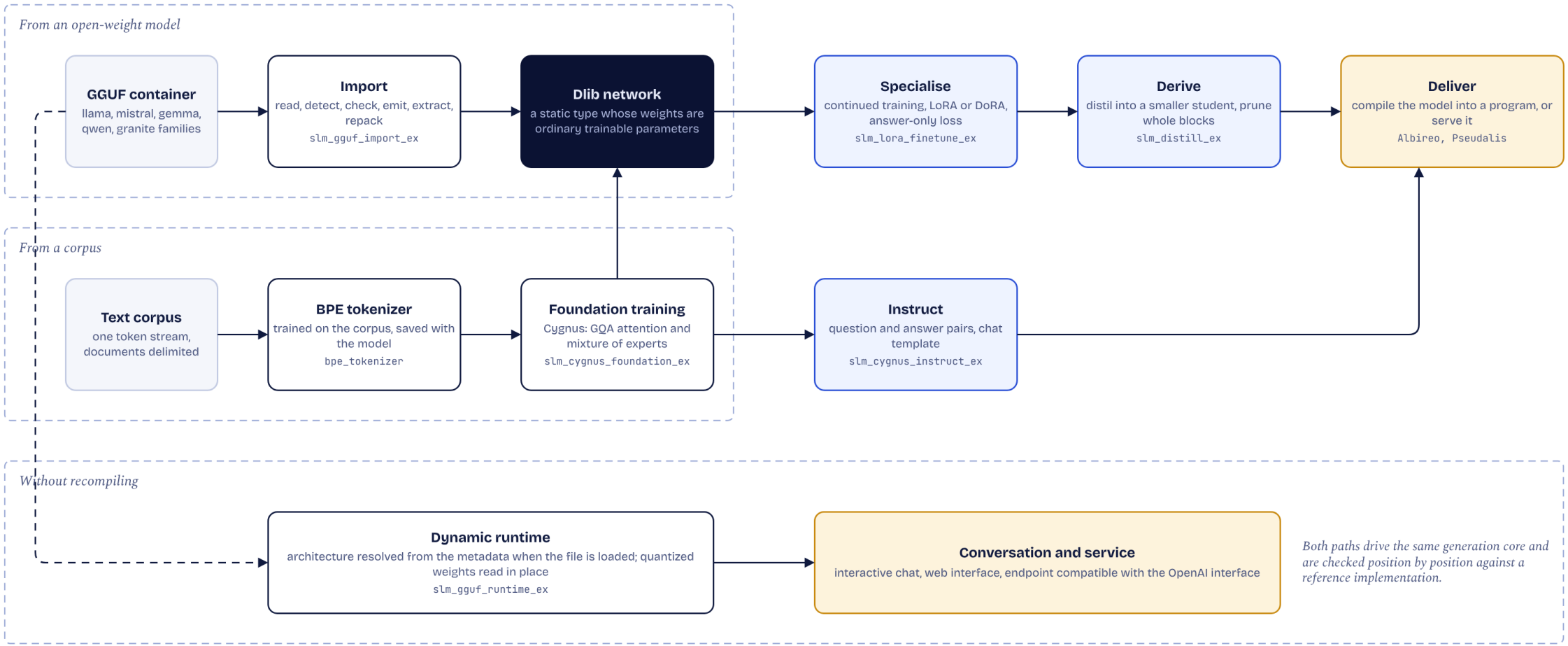
What returns

Fourteen contributions merged into Dlib between August 2024 and October 2025.

What it enables

Albireo and Pseudalis are built on the mechanisms the fork has stabilised.

One library for the whole life of a model



The architecture, layer by layer



- **Usable alone**

  A tokenizer without the runtime, the vision tower without a language model.

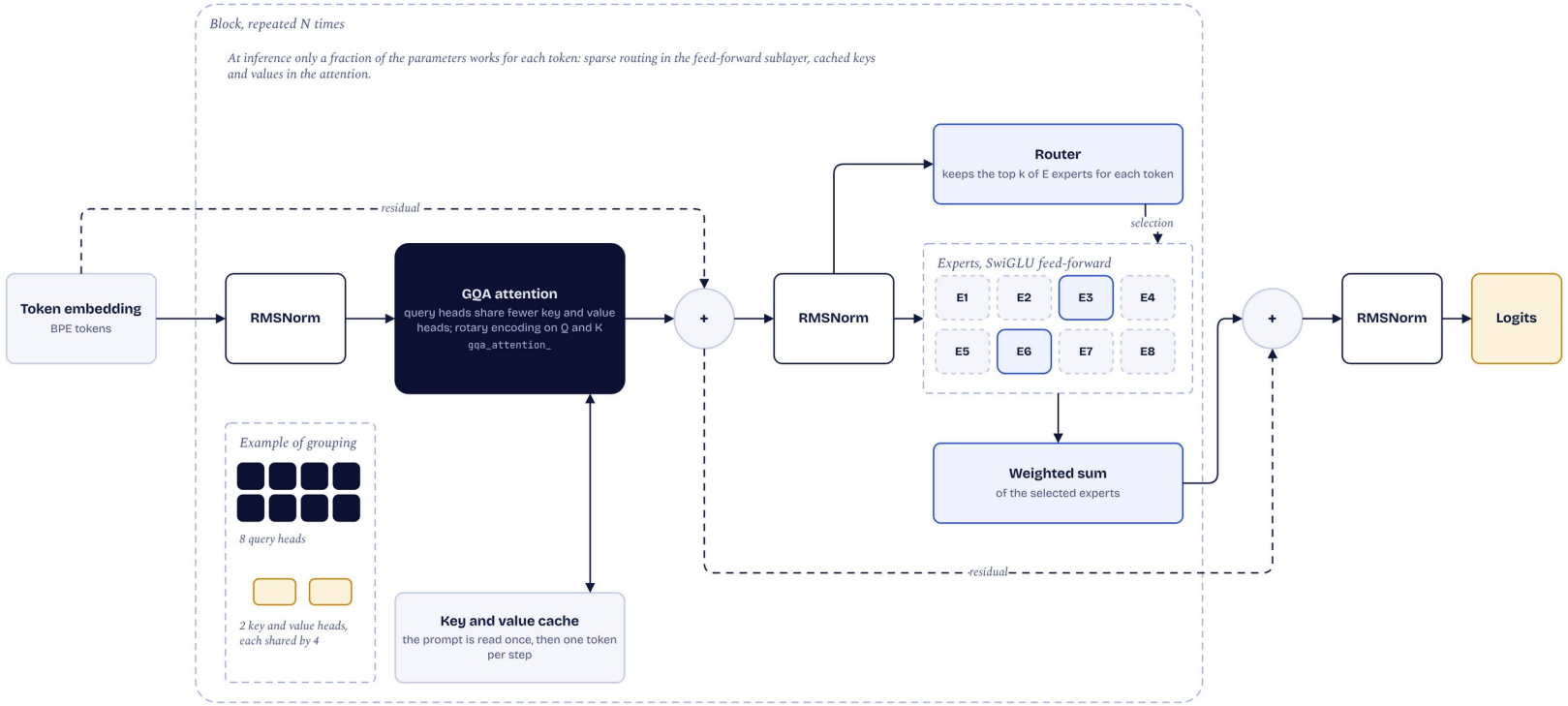
- **Below the network**

  Memory extension and device routing need no change to a model.

- **No Python**

  One language from the tensor operations to the installer.

Cygnus, the target architecture



- **Grouped query attention**

  Query heads share fewer key and value heads, with rotary position encoding.

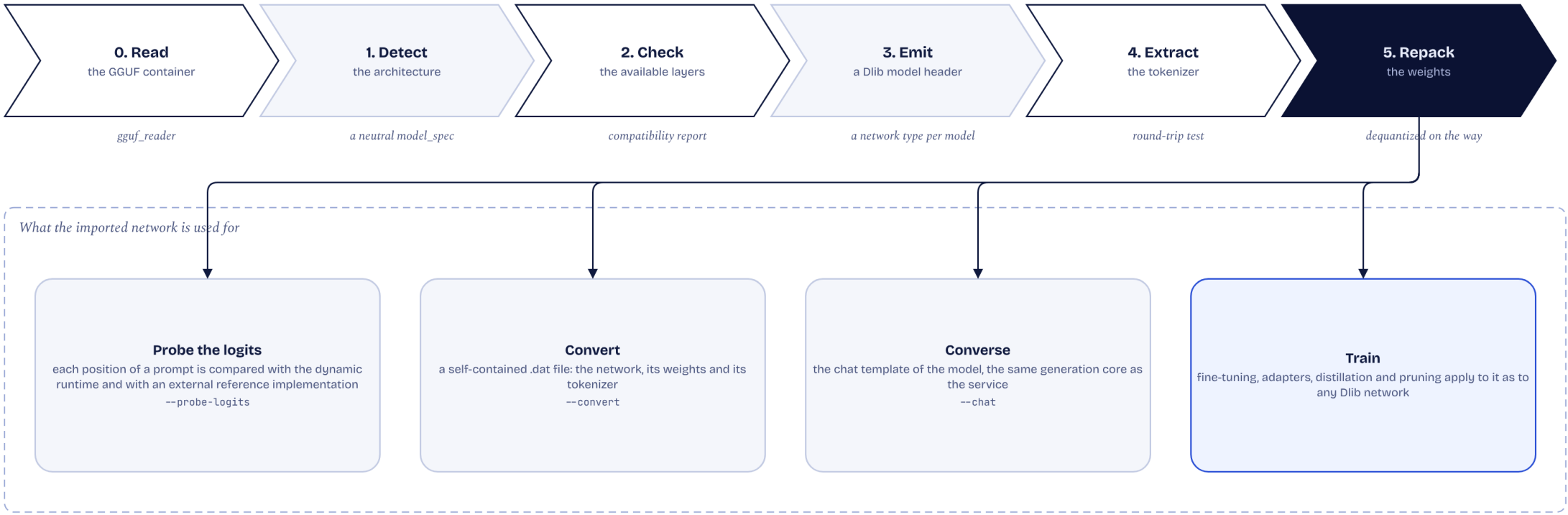
- **Mixture of experts**

  A router keeps the top k experts for each token.

- **Key and value cache**

  The prompt is read once, then one token per step.

Open-weight models, imported and verified



- **Twelve architectures**

  llama, mistral, gemma, qwen and granite families, dense and mixture of experts.

- **Verified**

  Each position of a prompt is compared with a reference implementation.

- **Trainable**

  The imported network is an ordinary Dlib network.

Specialise and derive a model

LoRA and DoRA

Adapters on the attention and feed-forward paths, merged into the weights afterwards.

`slm_lora_finetune_ex`

Distillation

A student of another architecture learns from the logits of a teacher.

`slm_distill_ex`

Depth pruning

Whole blocks of a published model are removed, and the rest repaired by distillation.

`slm_prune_distill`

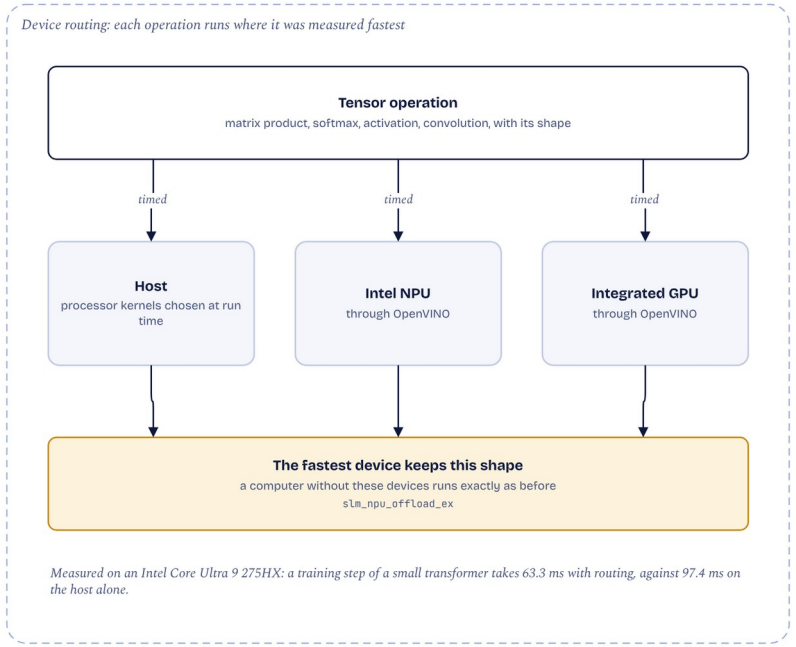
Embeddings and search

A folder of documents is indexed and answers a question with the passage that matches it.

`slm_embed_ex`

Memory and devices, handled below the network





17,506 MiB

working set trained on a card of 12,288 MiB

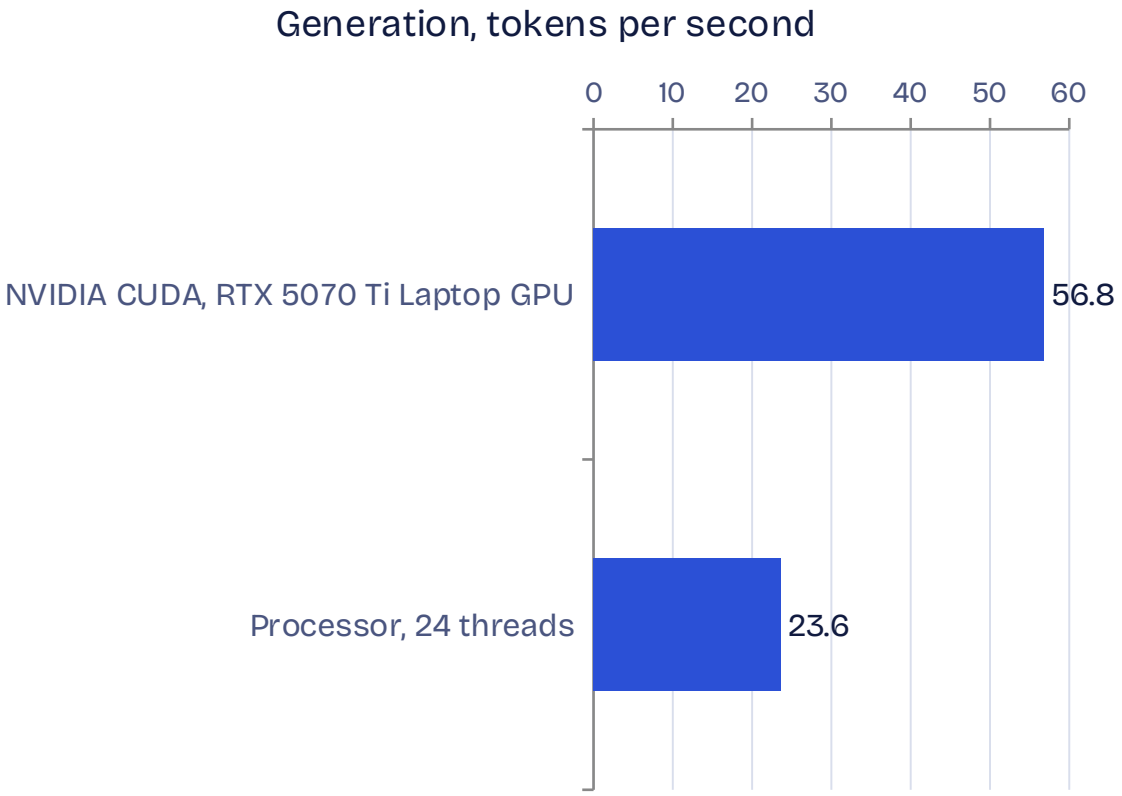
1.4×

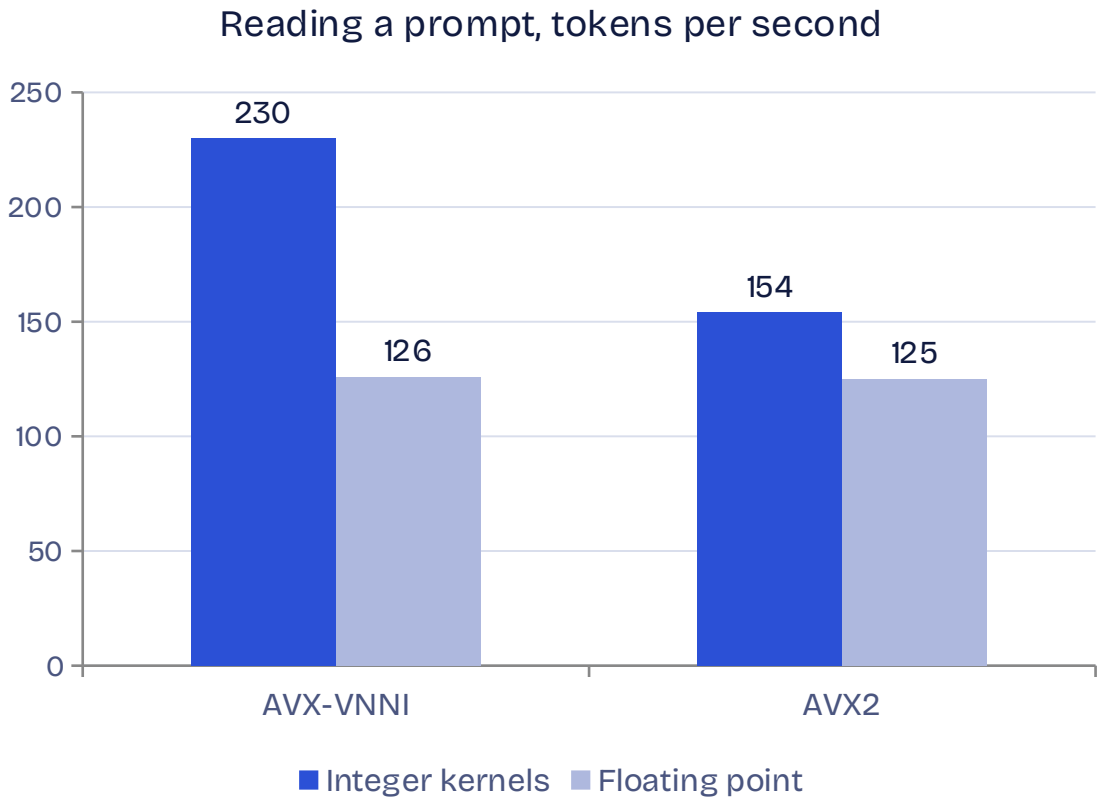
cost of training against a fully resident run

63.3 ms

training step with routing, against 97.4 ms on the host

Measured on ordinary hardware





Left: Qwen3-VL 4B in Q4_K_M. Right: prompt of 512 tokens, Qwen3 0.6B, two cores.

24 examples, one per mechanism

First models

A character-level model, then the canonical transformer.

Attention variants

Grouped queries, adaptive computation time, fused attention.

The Cygnus series

Foundation training, then instruction tuning and chat.

Open-weight models

GGUF import and runtime, LoRA, distillation, embeddings.

Vision

Vision tower, CIFAR-10 classifier, Barlow Twins.

Memory and devices

A network larger than the card, NPU routing.

Reasoning and applications

Sudoku and ARC-AGI, anonymization, compression.

`examples/slm_*`

Documentation that always stays up to date.

From the library to the products



Albireo

Language models that run on your own computer

56.8 tokens per second on a laptop GPU

22 models in the signed catalog

5 engines: CPU, Intel NPU and GPU, NVIDIA



Pseudalis

Personal data protection for documents, on your computer

0.946 F1 of the detector, five languages

41,000 characters read per second

40 countries with verified numbers

What is being stabilised now

QLoRA

Keep the frozen base quantized in memory during fine-tuning.

Preference optimisation

DPO with the frozen reference taken from the same network.

Mixture of experts, static path

Make these models trainable and convertible like the others.

A model of the library

A small language model with a topology and a vocabulary chosen here.

In line with P16

The work pursues the objectives of P16, the France 2030 programme led by Inria for sovereign, open and interoperable AI libraries. Cydral Technology intends to seek its support.

Cydral Technology

Use the library, or contribute to its development

github.com/Cydral/Dlib-Transformer-extensions
cydral-technology.eu/en/research/library/
cydral-technology.eu/en/contact/

Boost Software License 1.0. © 2026 Cydral Technology, France.