TORNADO VM

GPU-Powered AI Inference for Java

Pure Java. No JNI. No Python.
Just add a Maven dependency.

THE FULL JAVA AI STACK

Every layer is pure Java, production-ready, and Maven-installable

Your Application
Quarkus · Spring Boot · Microservices · CLI

LangChain4j / Quarkus-LangChain4j
RAG · Agents · Memory · Tool Calling · Streaming

GPULlama3.java
Llama 3 · Mistral · Qwen 3 · Granite · Phi-3 · Devstral — FP16 / Q8 / Q4

TornadoVM
JIT compiles Java → OpenCL / PTX / SPIR-V / Metal

GPU Hardware
NVIDIA · AMD · Intel · Apple Silicon

5×	270×	7+	24+
GPU speedup on 8B models	vs sequential Java CPU	model families	TornadoVM releases

MANCHESTER 1824
The University of Manchester

tornadovm.org · gpullama3.dev

Seamless Integration

Drop-in GPU inference for the Java ecosystem you already know

LANGCHAIN4J - v1.15.0 / Maven + Gradle

```
// Maven (pom.xml)
<dependency>
  <groupId>dev.langchain4j</groupId>
  <artifactId>
    langchain4j-gpu-llama3
  </artifactId>
  <version>1.15.0-beta25</version>
</dependency>

// Gradle (build.gradle)
implementation 'dev.langchain4j:
  langchain4j-gpu-llama3:1.15.0-beta25'

// Use like any ChatLanguageModel
GPULlama3ChatModel.builder()
  .modelPath(Paths.get("llama3.gguf"))
  .onGPU(true).build();
```

QUARKUS - cloud-native AI (v1.10.0)

```
// Maven (pom.xml)
<dependency>
  <groupId>
    io.quarkiverse.langchain4j
  </groupId>
  <artifactId>
    quarkus-langchain4j-gpu-llama3
  </artifactId>
</dependency>
```

MATURITY — production-ready since 2025

- ☑ **Official Integrations**
  First-class provider in LangChain4j (v1.15.0) and Quarkus-LangChain4j (v1.10.0). Maintained by core team.
- ☑ **Multi-Backend GPUs**
  OpenCL (NVIDIA, AMD, Intel), PTX/CUDA (native NVIDIA), SPIR-V (Level Zero), Metal (Apple M1–M5).
- ☑ **Enterprise Java Stack**
  JDK 21–25. OpenJDK, GraalVM, Temurin, Zulu, SapMachine, Mandrel. Maven Central + SDKMAN!
- ☑ **7+ Model Families**
  Llama 3, Mistral, Devstral, Qwen 2.5/3, Phi-3, IBM Granite 3.2/4.0. GGUF with FP16, Q8, Q4.
- ☑ **Battle-Tested at Scale**
  24+ releases, 10K+ commits, 30+ contributors worldwide. Nightly CI across all backends and JDKs.
- ☑ **Real-World Impact**
  ESA GAIA space mission. 7+ EU projects (AERO, TANGO, P2CODE, Encrypt, ELEGANT).

Same LangChain4j API. Same tooling. GPU is transparent.
Works with RAG, agents, memory, streaming, tool calling.

Resources

- **TornadoVM**
  `github.com/beehive-lab/TornadoVM`
- **GPULlama3.java**
  `github.com/beehive-lab/GPULlama3.java`
- **Docker Image**
  `hub.docker.com/r/beehivelab/gpullama3.java-nvidia-openjdk-opencl`
- **LangChain4j Demo**
  `github.com/beehive-lab/gpullama3-langchain4j-demo`
- **Quarkus Demo**
  `github.com/beehive-lab/gpullama3-quarkus-langchain4j-demo`
- **SDKMAN!**
  `sdkman.io/sdks/tornadovm`

Scan to get started

`gpullama3.dev`

Supported by:
EU Horizon 2020 & Horizon Europe
UK Research and Innovation