

Quarkus meets AI :

Build your own LLM-powered application



How can I help you today?

Brainstorm incentives

for a customer loyalty program in a small bookstore

Give me ideas

for what to do with my kids' art

Suggest fun activities

to do indoors with my high-energy dog

Show me a code snippet

of a website's sticky header

Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.



Zineb Bendhiba

- Open Source Engineer at **Red Hat**
- **Apache Camel** PMC
- International Speaker
- 15+ years professional software development experience
- Speak English, French, Moroccan Darija, Arabic
- **University of Cadi Ayyad** Alumni
- <https://zinebbendhiba.com>
- Twitter: @ZinebBendhiba



Generative AI

ChatGPT

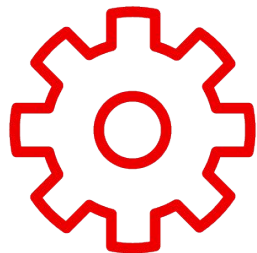
[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

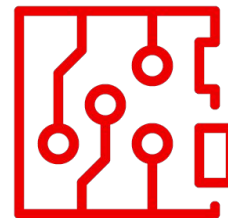
ChatGPT (Chat Generative Pre-trained Transformer) is a large language model-based chatbot developed by OpenAI and launched on November 30, 2022, that enables users to refine and steer a conversation towards a desired length, format, style, level of detail, and language. Successive prompts and replies, known as [prompt engineering](#), are considered at each conversation stage as a context.^[2]



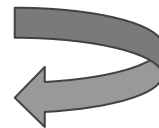
Large Language Model - 101



Turn **Base Model** into a
Chat Assistant Model



Evaluate, Deploy,
Monitor, Re-train



Model Pre-training (self-supervised)

- Prepare/configure the **billion parameters***
- Collect/tokenize **trillions of data***
- Let the model **self-train** on that data
- Model learns to predict the **next word**
 - E.g. "once upon a" \Rightarrow "time" (99%)
 - But it can also make up stuff!
- Requires **\$M(MM)** in GPU/processing power and days to weeks of training

Model Fine-tuning (supervised)

- Curate **hundreds** of high quality chats (**Q&As**)
- Let the model **retrain** on the new data set
- Model learns to **mimic the chat behaviour**
 - Understands instructions
 - Using the pre-trained knowledge!
- **[Retrain to make it Helpful, Honest, Harmless]**
 - Reinforcement Learning from Human Feedback (RHFL)
- Takes a day of processing for each cycle



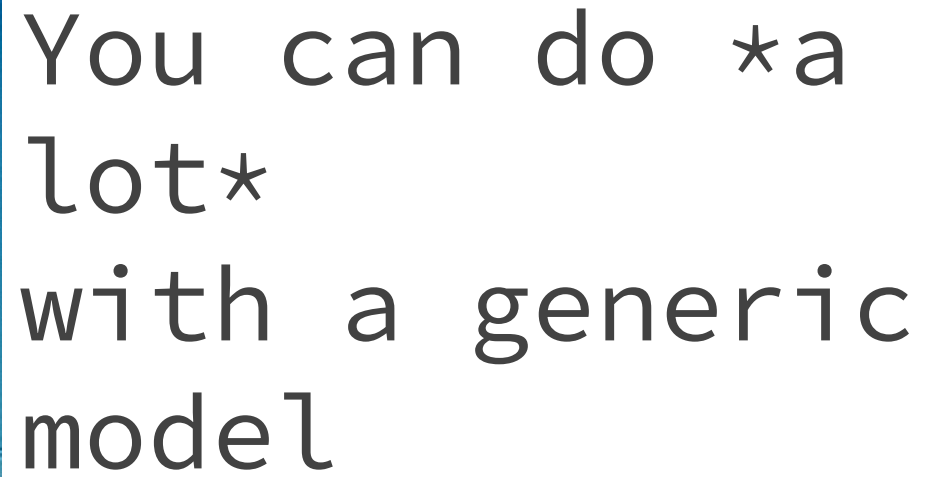
This is
Data Science
stuff!

What about the Developers?

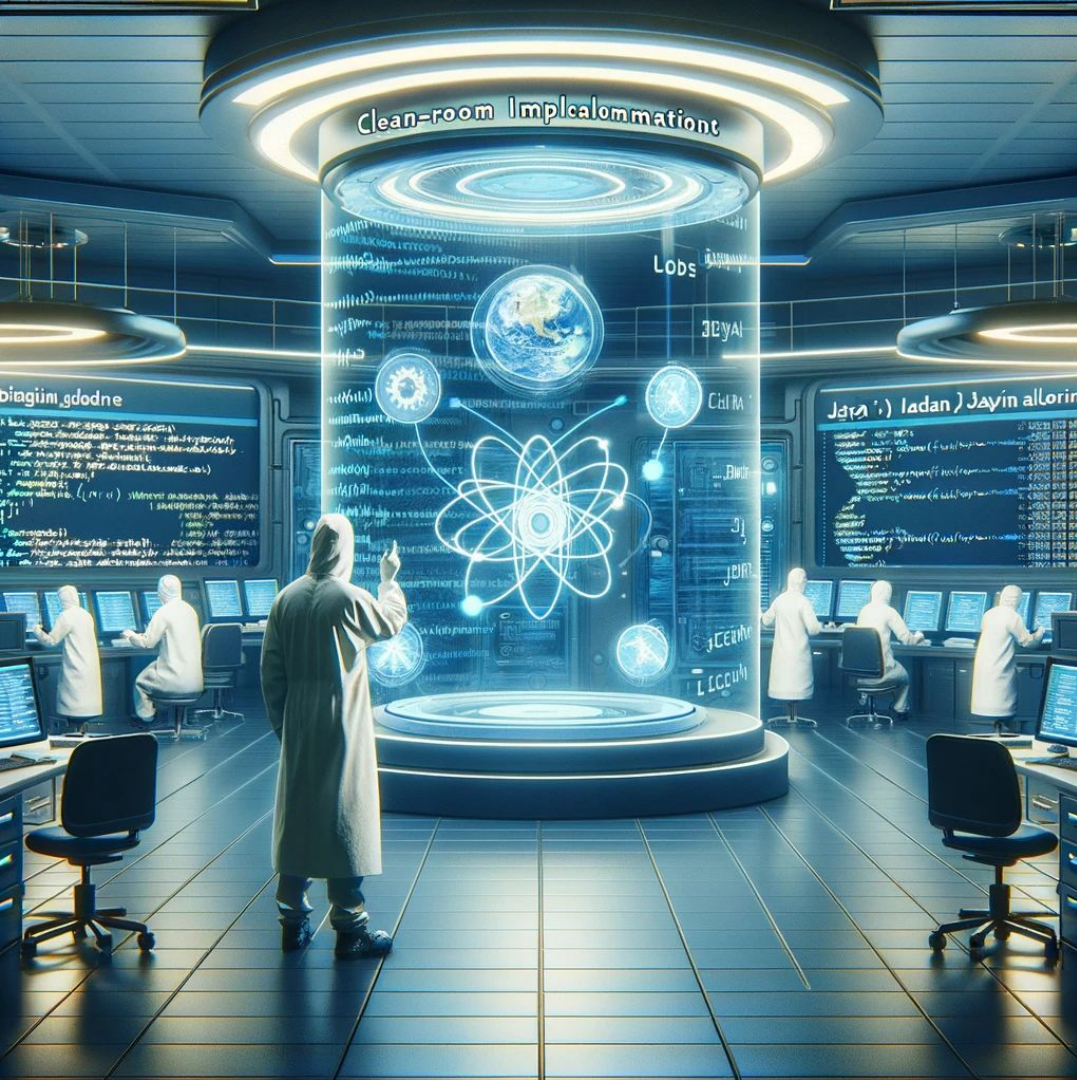


The "Model" is
only part of
your
Architecture

It's just another API



...and enough context
(docs, emails, other data)



LangChain



Java™

LangChain4j 

Components of LangChain4j

Chains

Tools

AI Services

New!

Basics

Image
Models

Language
Models

Prompt
Templates

Output
Parsers

Memory

RAG

Document
Loaders

Document
Splitters

Embedding
Models

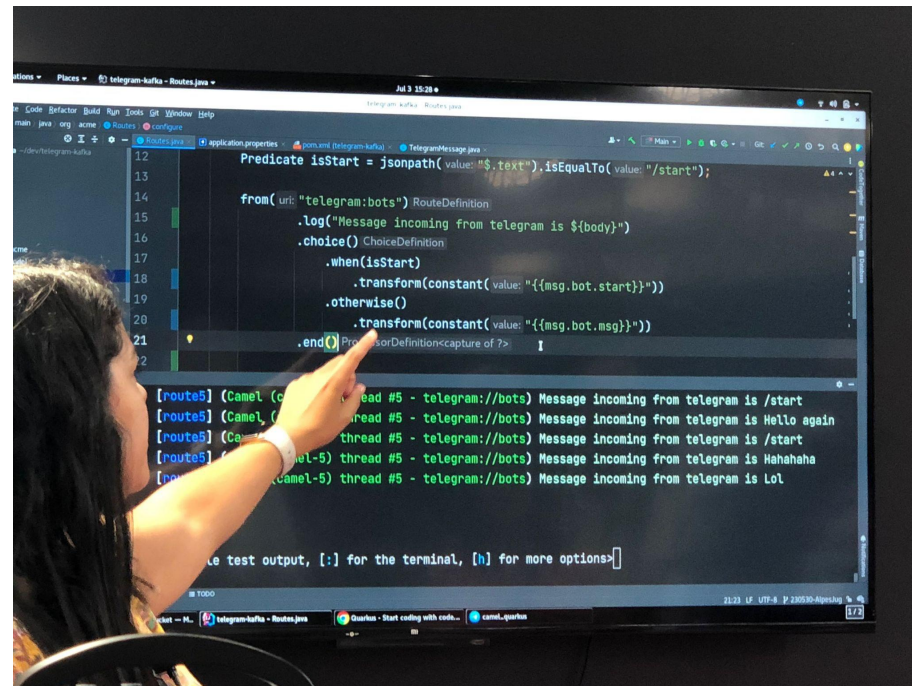
Embedding
Stores

Why LangChain4j with Quarkus?

- Programming Model
 - Seamless integration with CDI
 - Simpler declarative AI Services (`@RegisterAiService`)
- Developer Joy
 - 'quarkus dev' enables iterative testing and Prompt tuning.
 - Dev UI allows to view AI services, tools, config, add/search embeddings, test prompts, generate images
- Performance Enhancements
 - Optimized Quarkus REST/JSON libs
 - Reduced library footprint
- Build time
 - Build time warnings
 - Compile to native with GraalVM
- Production Enhancements
 - Unified Configuration
 - Metrics, Tracing, Auditing



Demo



Links

- [Quarkus Langchain 4j documentation](#)
- [Quarkus Langchain4j samples](#)
- [Apache Camel project](#)
- [Quarkus project](#)
- [Langchain4j project](#)
- [Demo](#)
- [My Apache Camel talk af Blabla conf 2021](#)
- [WIP: Camel Langhcaain4j component](#)

Thank you



Zineb Bendhiba (@ZinebBendhiba)