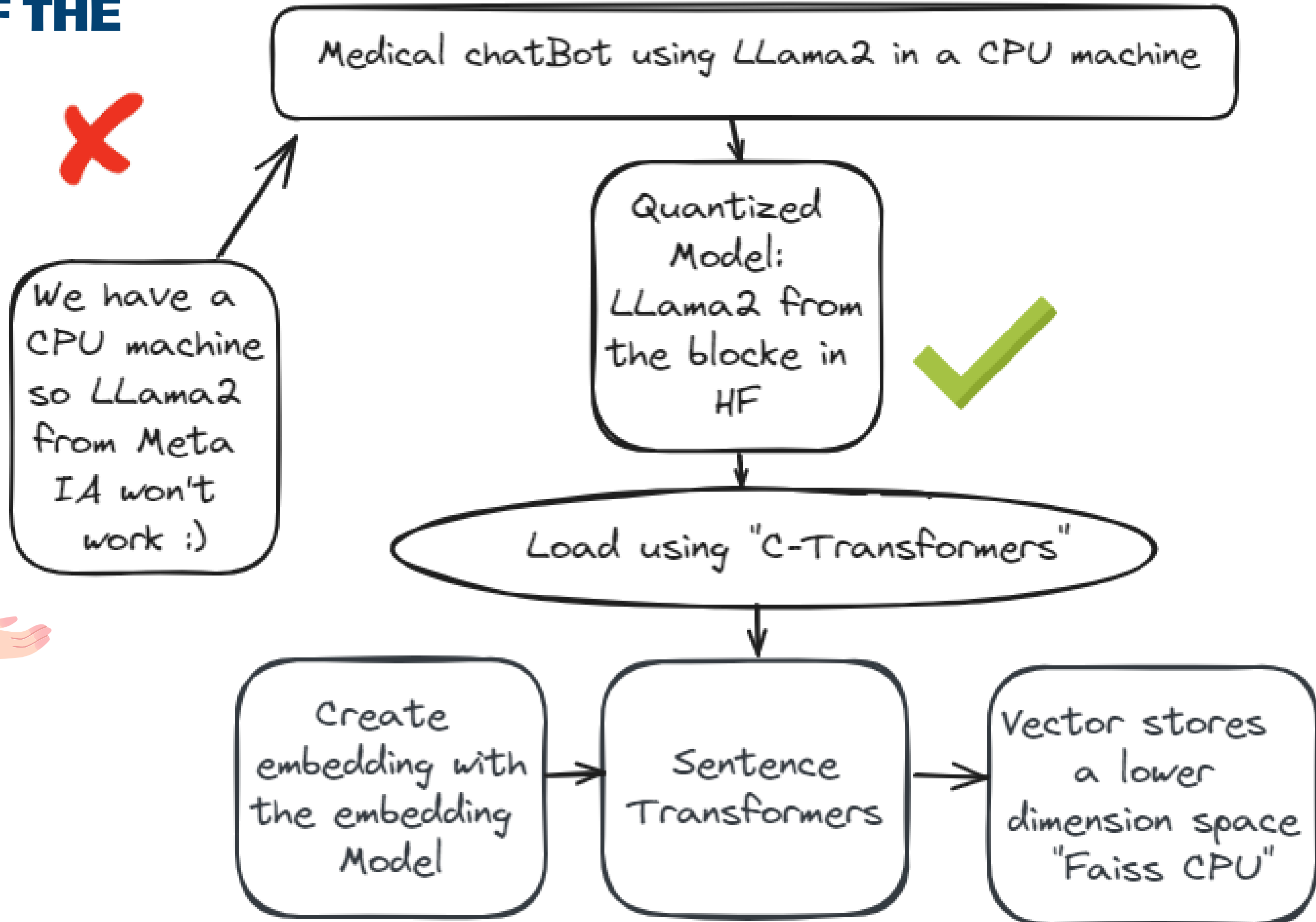
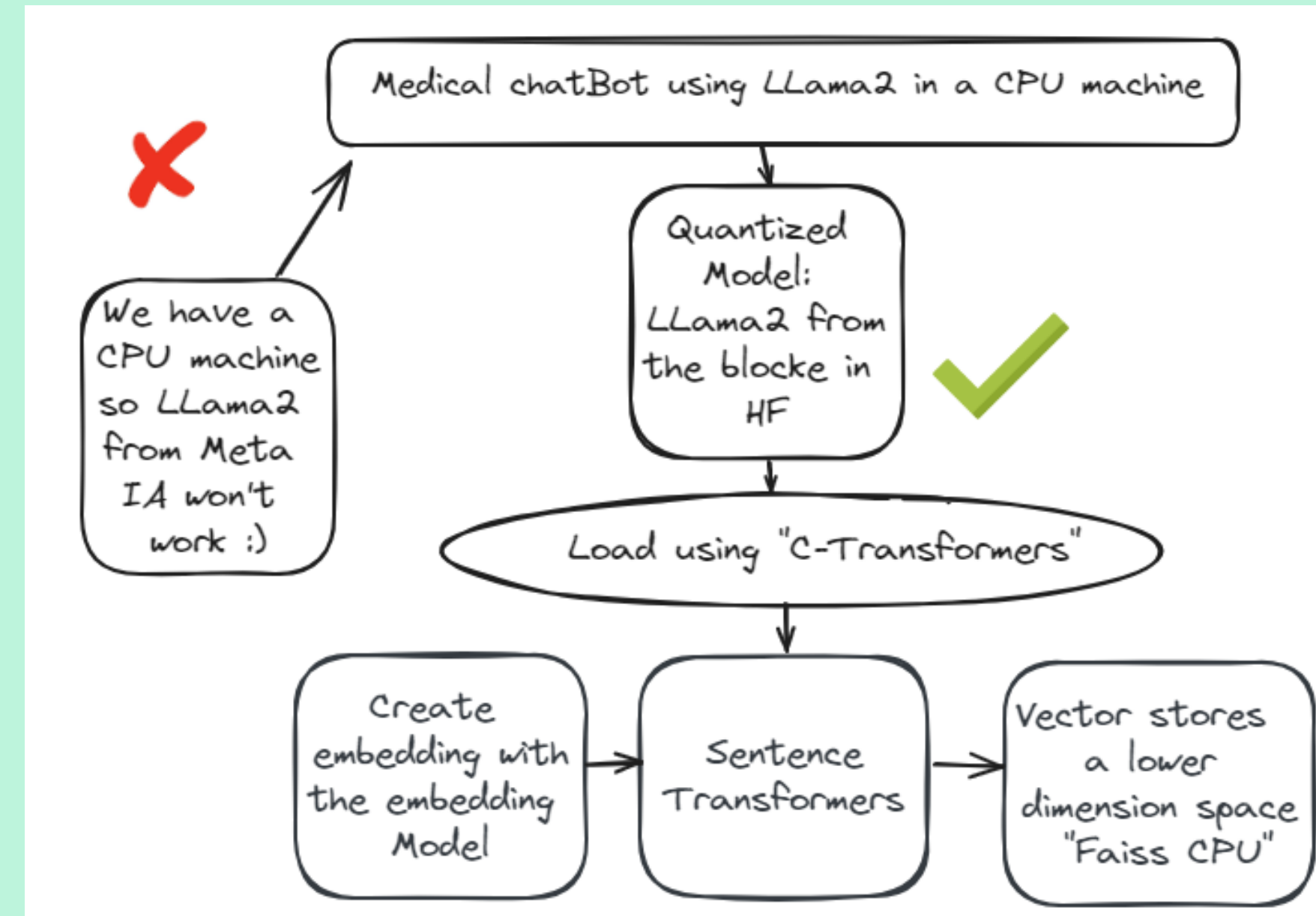


# SCHEMA OF THE WORKFLOW OF THE PROJECT



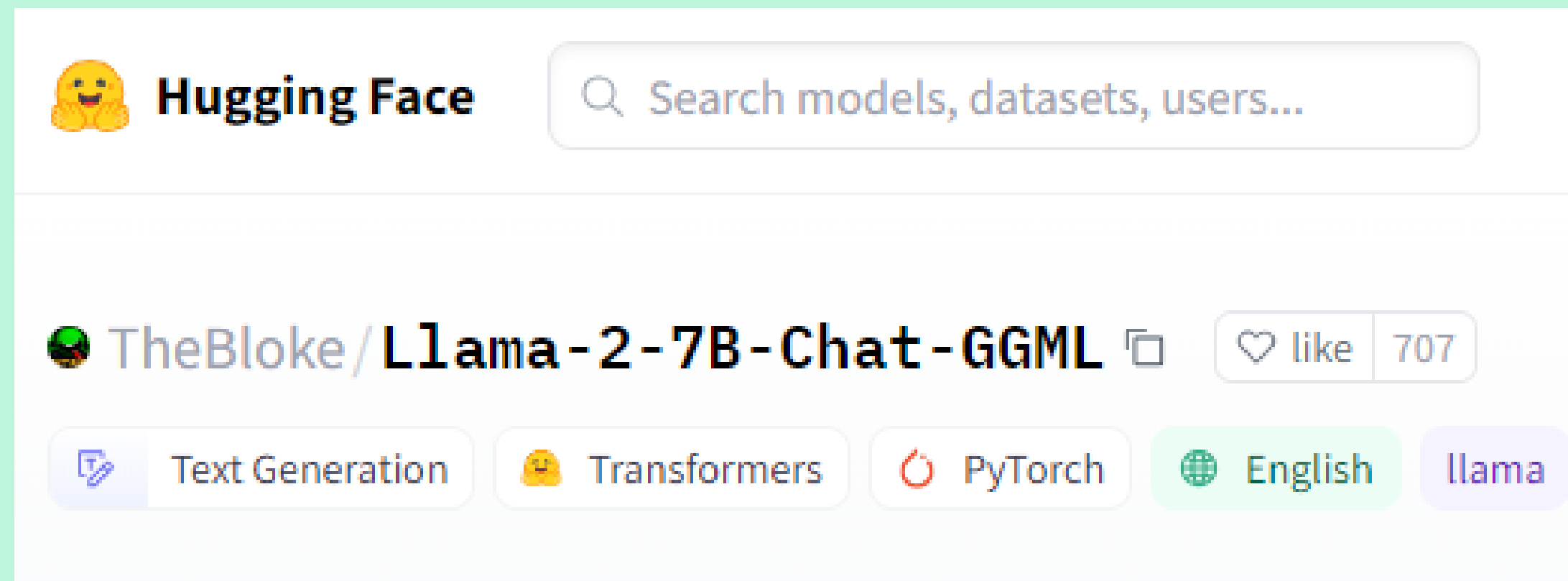
# STEPS

- **Objective:** Make advanced AI models more accessible on CPU setups.
- **Challenge:** LLama2 model by Meta AI initially incompatible with CPUs due to GPU optimization.
- **Solution:** Quantized version of LLama2 model, reducing precision without compromising performance, available via Hugging Face model repository.
- **Implementation:** Utilized 'C-Transformers,' a CPU-optimized library to efficiently run the quantized model on a CPU-based machine.
- **Data Processing:** 'Embedding model' converts raw text data into numerical embeddings capturing semantic meaning.
- **'Sentence Transformers'** method refines embeddings for natural language context and nuance preservation.
- **'Faiss CPU'** used for storing and managing embeddings in a lower-dimensional space, enhancing chatbot speed and efficiency, even on less powerful CPUs.



# QUANTIZATION

- 'Quantization' is a technique that reduces the precision of the model's computations, making it lighter and faster, without significantly compromising its performance.



<https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML>

# C-TRANSFORMERS



- C-Transformers is a library or tool designed to optimize the loading and execution of deep learning models, including transformer models, on CPU (Central Processing Unit) machines.
- Its primary purpose is to enable the deployment and inference of deep learning models on hardware that lacks the specialized hardware acceleration (e.g., GPUs) that is often used for AI tasks.



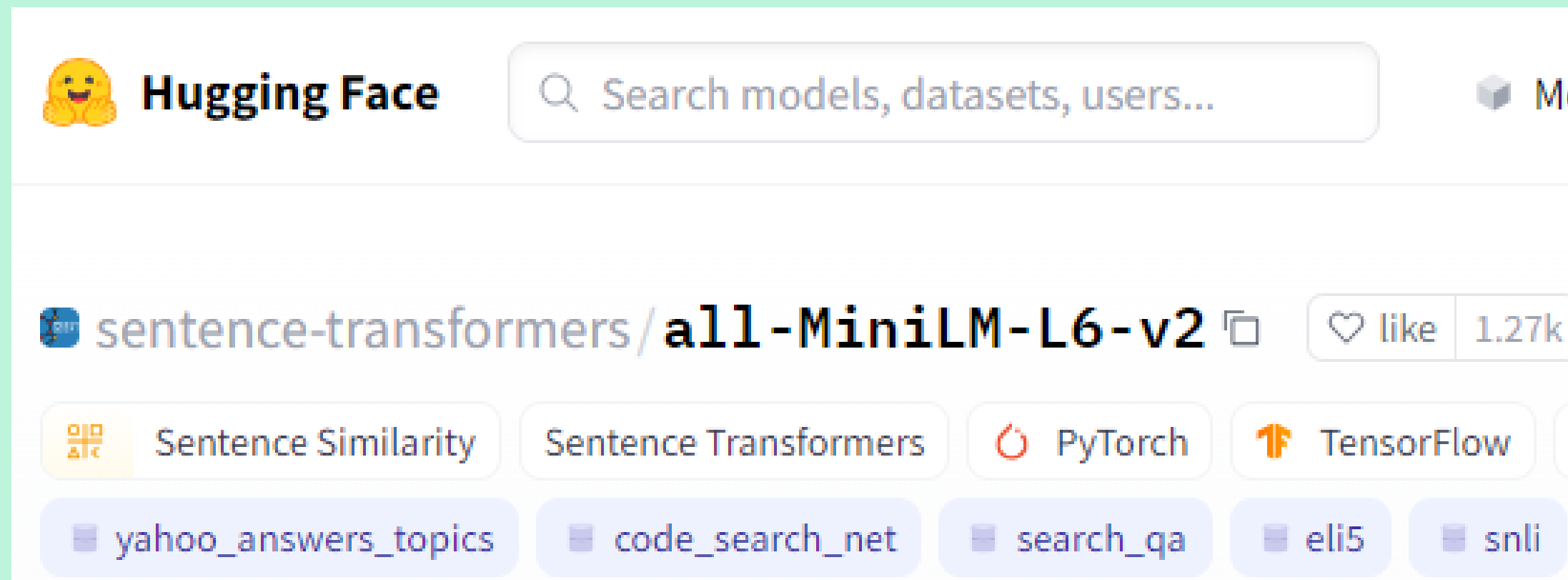
## Installation and Setup

- Install the Python package with “**pip install ctransformers**”

# SENTENCE TRANSFORMERS

- Is a technique or approach used to refine embeddings (numerical representations) of sentences or text passages.
- Enhance the quality and relevance of sentence embeddings, making them more suitable for various NLP tasks such as information retrieval, semantic similarity, and text classification.

**“pip install -U sentence-transformers”**




- It maps sentences & paragraphs to a 384 dimensional dense vector space



# FAISS CPU : FACEBOOK AI SIMILARITY SEARCH

- Faiss is a library for efficient similarity search and clustering of dense vectors.
- Faiss is written in C++ with complete wrappers for Python/numpy.
- FAISS includes techniques like vector quantization and CPU Optimization.
- Optimized for CPU-based computations. This makes it valuable for deployments on systems that lack GPU




 ANACONDA.ORG


Search Anaconda.org


pytorch / packages / faiss-cpu 1.7.4


A library for efficient similarity search and clustering of dense vectors.

Conda	Files	Labels	Badges
-------	-------	--------	--------

 License: MIT

 Home: <https://github.com/facebookresearch/faiss>

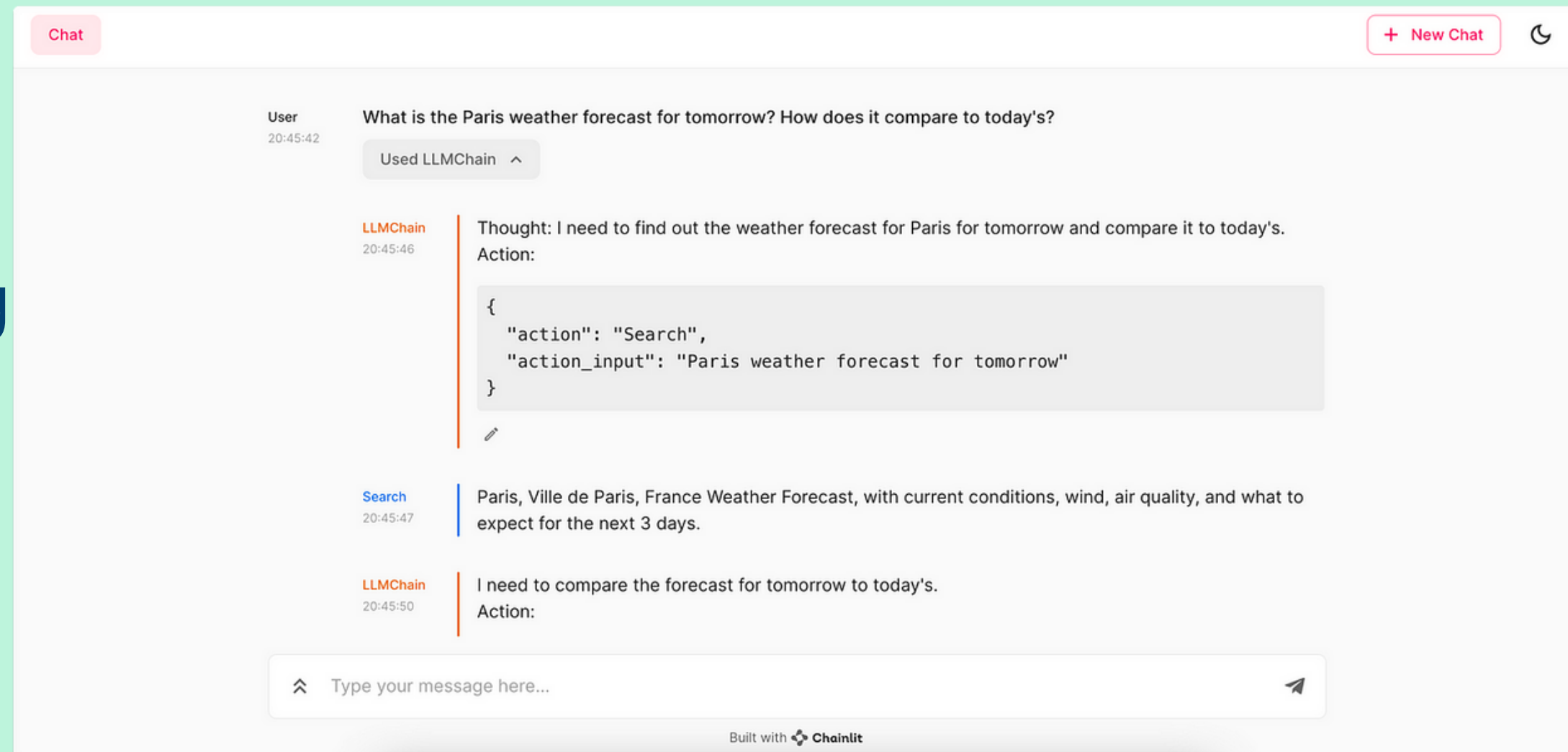
 1581382 total downloads

 Last upload: 13 minutes and a few seconds ago



# CHAINLIT LIBRARY

- Chainlit is a Python library that lets us build Chat Interfaces for Large Language Models in minutes.
  - \$ pip install chainlit
  - \$ chainlit hello
  - \$ chainlit run MedBot.py



# ARCHITECTURE

## HOW IS A RESPONSE GENERATED WHEN INTERACTING WITH THE CHATBOT?

The figure aside illustrates the mechanism relying behind the generation of a response according to a given prompt!

