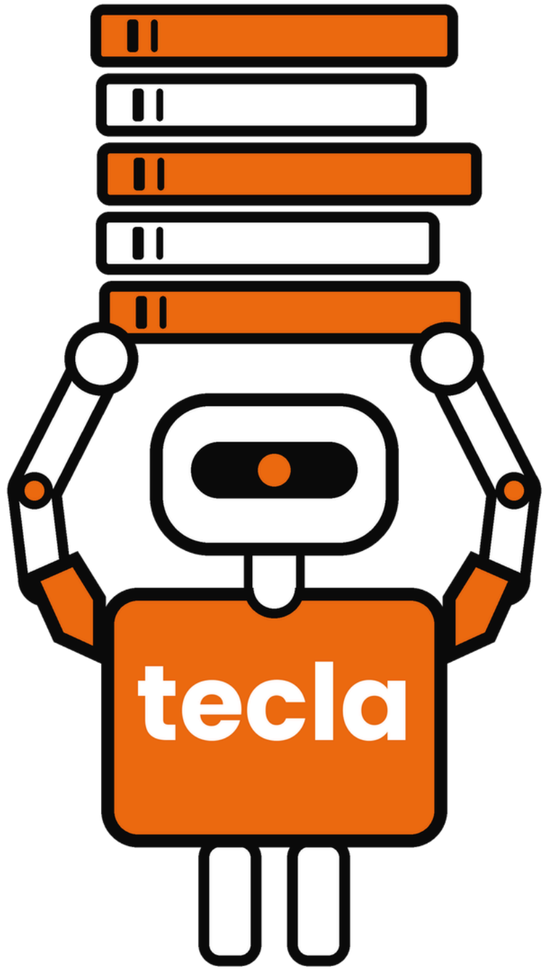


tecla

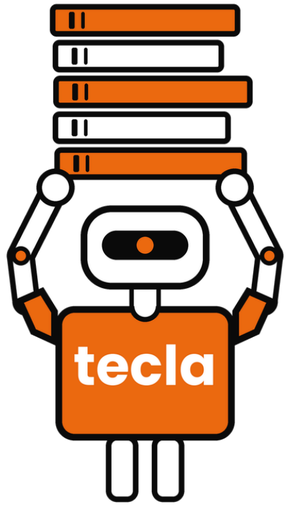
(by Silicon Team)

Gianluca Lorusso
Mohamed Dhia Rezgui
Farouk Fadel Brachemi
Nidhal Karchoud

TECHPARK
SÜDTIROL / ALTO ADIGE

TELMEKOM
The future is easy

tecla is based on a hybrid RAG+LLM architecture.



It runs entirely locally thanks to llama.cpp. This ensures **speed** and **data sovereignty**.

Answers are grounded in the files' content (PDF, markdown and images).
No web sources used.

Supported languages: EN, IT, DE.

Hybrid retrieval combines BGE-M3 semantic search, BM25 keyword matching, and Reciprocal Rank Fusion across 918 indexed manual sections, preserving product-specific evidence.

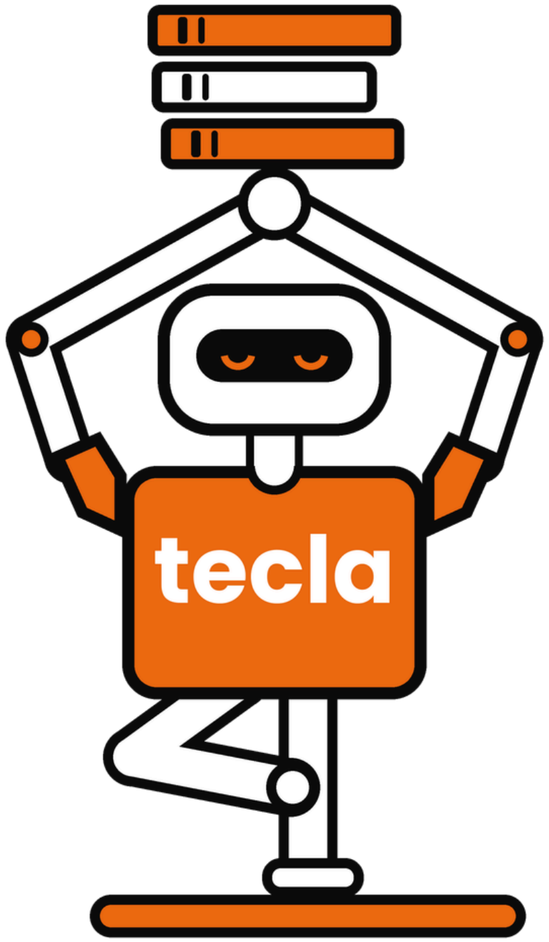
This hybrid approach pays dividends when eventually extending the dataset.

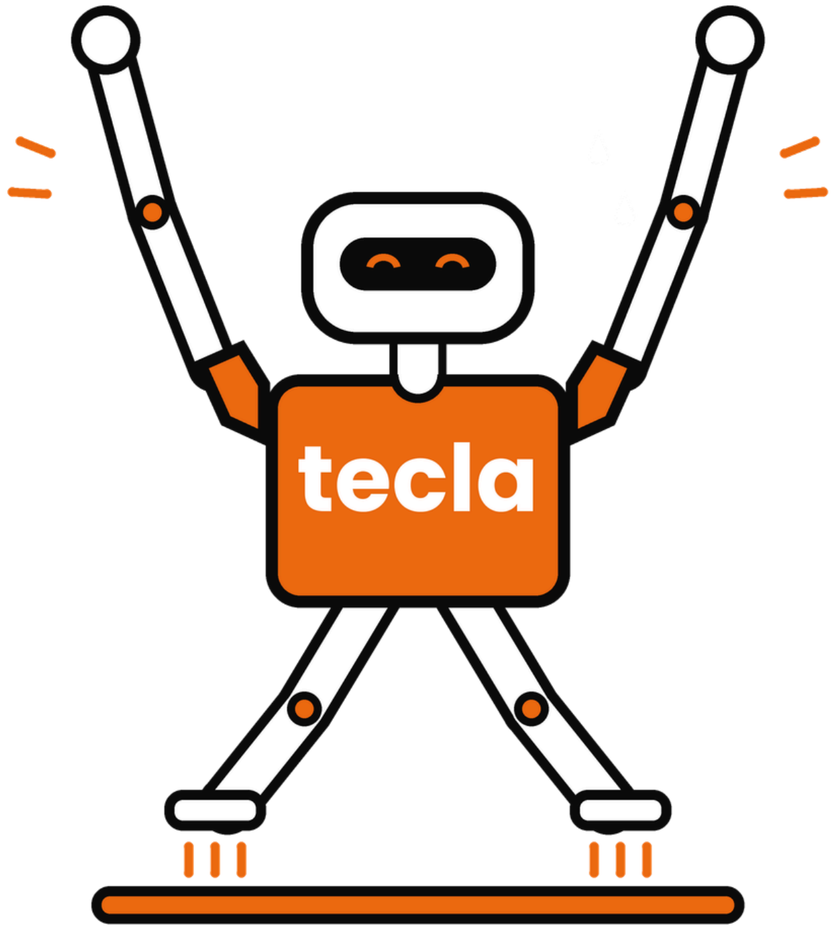
- **Multiple models were compared**
- **For instance, out of 63 prompts in 3 languages (189 total prompts) 188/189 gemma4-e4b produced a grounded accurate result.**
- **Cuda GPU acceleration is also supported to speed up inference.**
- **Ultimately any model compatible with llama.cpp can be specified in the config.**

tecla can do **yoga**.

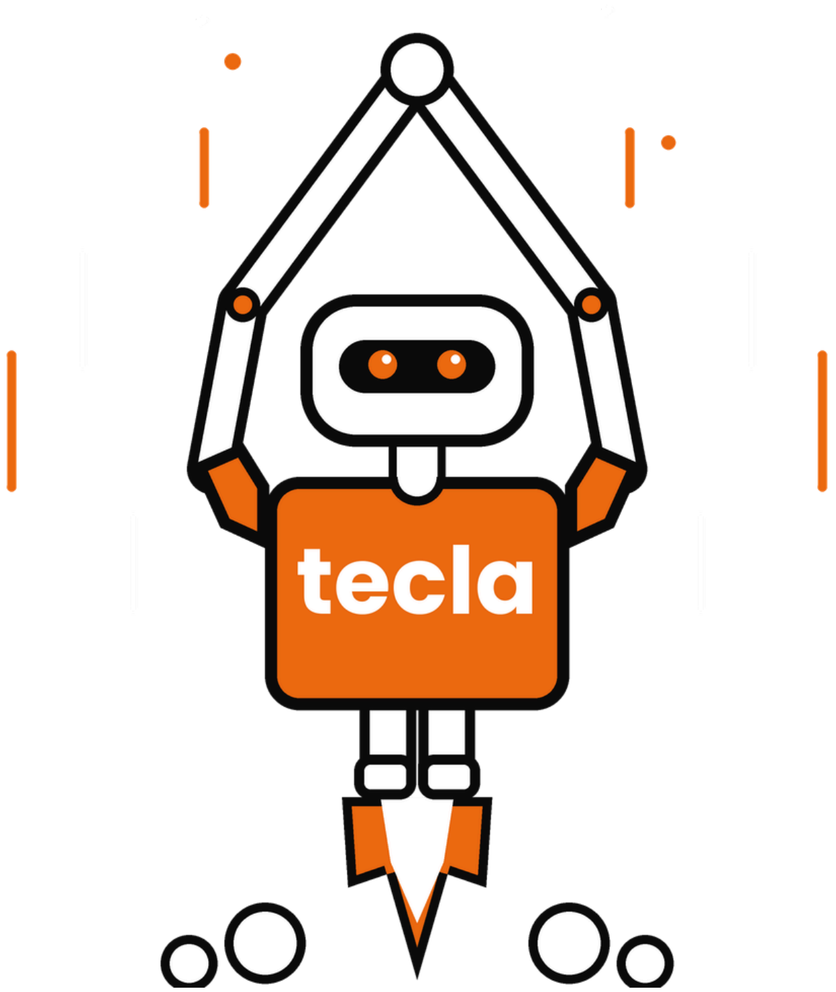
It’s flexible, customizable and scalable with docker.

The most relevant parameters, including the models themselves, can easily be changed in config files.





DEMO TIME



Thank You

Gianluca Lorusso
Mohamed Dhia Rezgui
Farouk Fadel Brachemi
Nidhal Karchoud

TECHPARK
SÜDTIROL / ALTO ADIGE

TELMEKOM
The future is easy