

LLM in the browser



Why in the browser?

- The app works out of the box.
- There is no need for a backend.
- **It scales infinitely.**

Hugging Face

transformer.js

- question-answering
- text2text generation
- text generation

Natural Language Processing



Text Classification



Token Classification



Table Question Answering



Question Answering



Zero-Shot Classification



Translation



Summarization



Feature Extraction



Text Generation



Text2Text Generation



Fill-Mask



Sentence Similarity

Curriculum Vitae Chatbot

text2text generation

Model

- Xenova/LaMini-Flan-T5-783M
- Ca. 800 MB
- No tokenizer

Prompt:

`Answer this question "\${question}"
given this context "\${this.context}."`

Chat with my AI-Powered Agent (Alpha)

Agent

Hello, I'm Stefano's AI-Powered agent. How can I assist you today?

You

Where does he work?

Agent

He works as a Senior Software Engineer at Schaltstelle GmbH in Bern, Switzerland.

You

What does he study?

Agent

He studies Artificial Intelligence at Berner Fachhochschule BFH.

Type your question

Send

question-answering

Model

- Xenova/distilbert-base-cased-distilled-squad
- Incl. tokenizer 200 MB

No prompt:

- `this.qnaPipeline(question, mostSimilarText?.text)`

The screenshot shows a web application with a dark blue header containing navigation links: Home, Projects, Talks, and Courses. The main content area is white and titled "Chat with my AI-Powered Agent (Alpha)". The chat interface displays three messages:

- Agent:** Hello, I'm Stefano's AI-Powered agent. How can I assist you today?
- You:** Where does he work?
- Agent:** Schaltstelle GmbH
- You:** What does he study?
- Agent:** Artificial Intelligence

At the bottom, there is a text input field with the placeholder "Type your question" and a blue "Send" button. A red banner at the very bottom of the page contains a chat icon and the text "Chat with my AI-Powered Agent (Alpha)".

Let's check the code

End

Conclusion

- Small models are weak
- Bad for a curriculum vitae chatbot
- Good for Progressive Web App
 - Install on devices

Next

- Use a backend like Firebase Functions

Important

- Soon: Built-in LLM in browser or OS