LLM in the browser

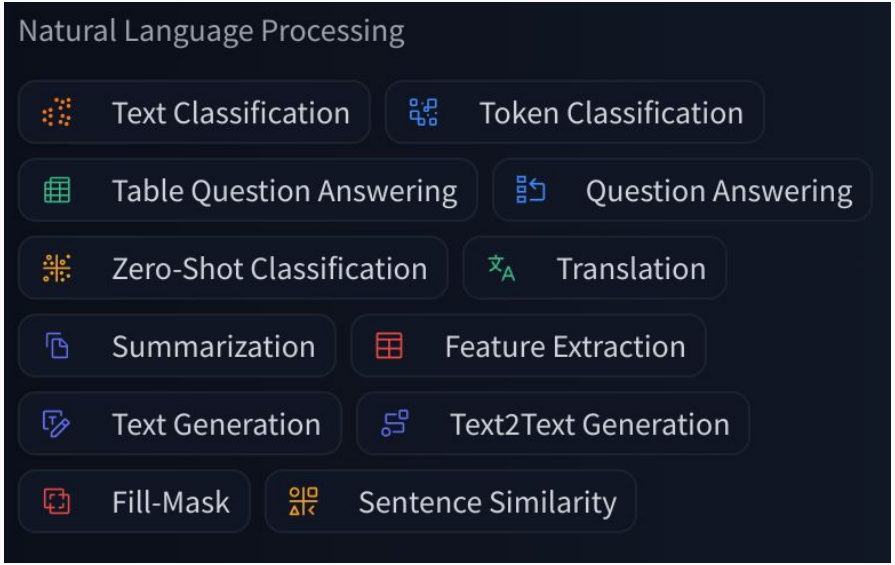
Why in the browser?

- 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



Curriculum Vitae Chatbot

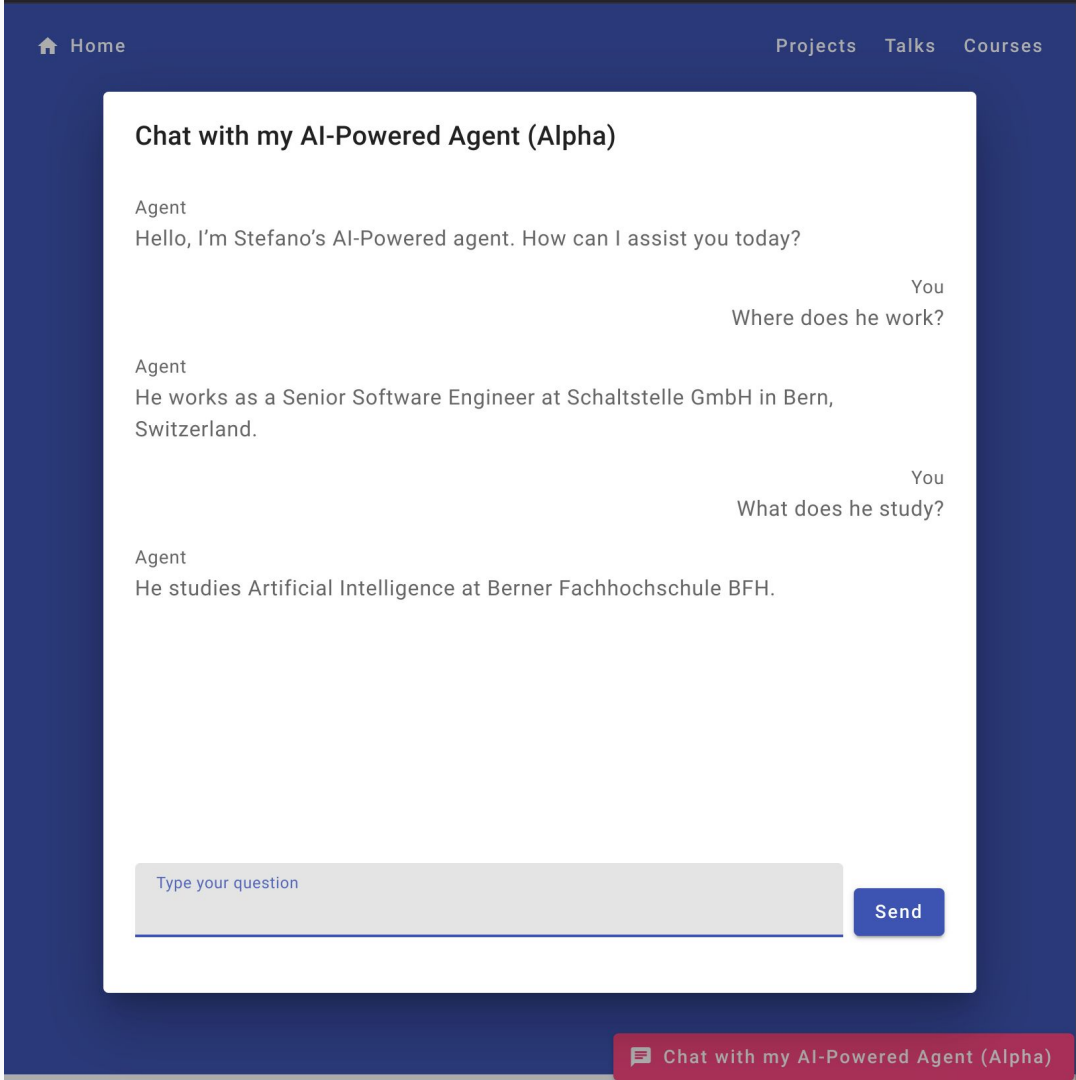
text2text generation

Model

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

Prompt:

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



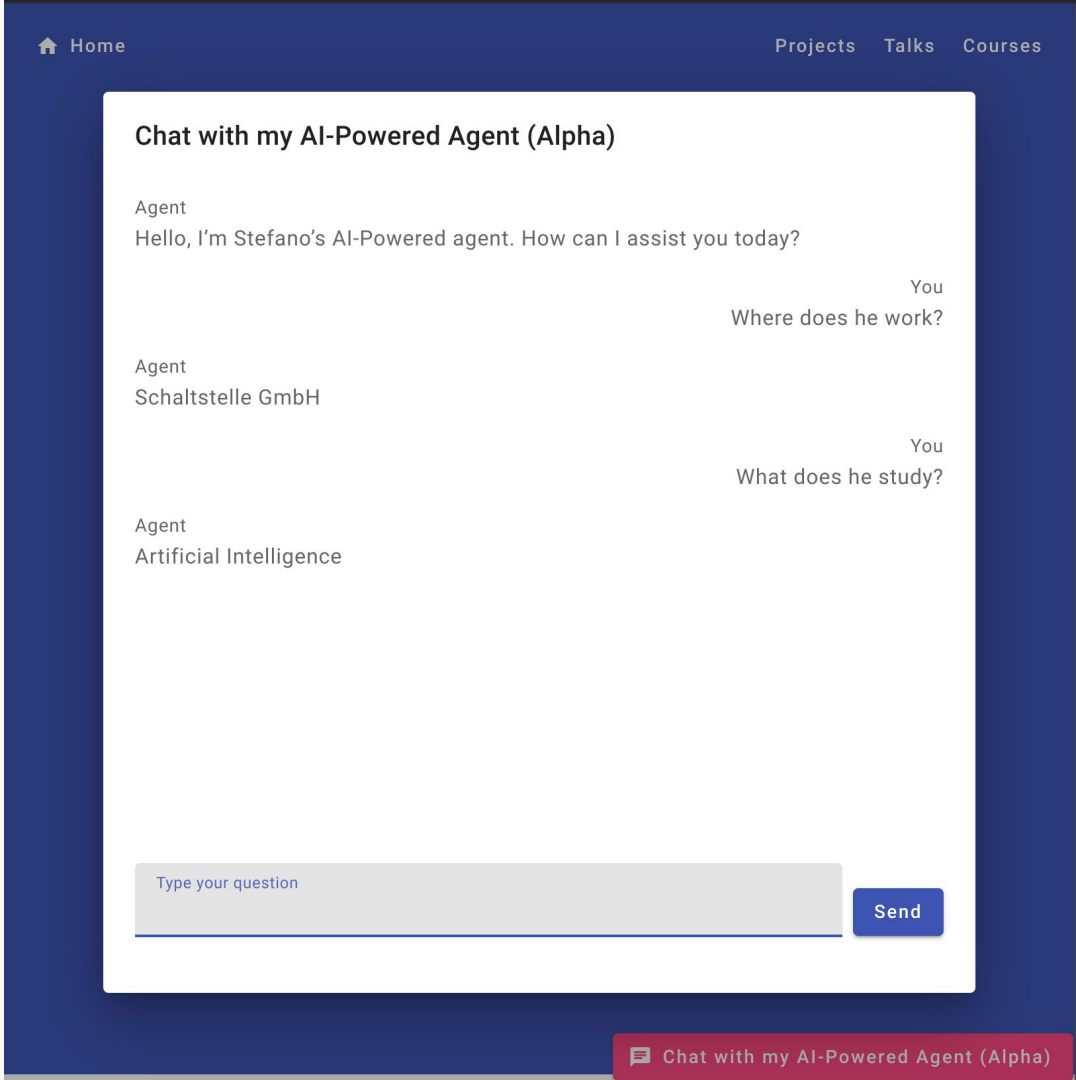
question-answering

Model

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

No prompt:

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



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