Recherche:

* tokenization: <https://www.youtube.com/watch?v=D8j1c4NJRfo>
* batch: <https://www.youtube.com/watch?v=T0VSn-XkjGM>, <https://www.youtube.com/watch?v=cjBn4ATxZoQ>
* decoder: https://www.youtube.com/watch?v=Qo\_0PGusJfU
* pytorch: <https://www.youtube.com/watch?v=ORMx45xqWkA>, <https://www.youtube.com/watch?v=r2JKV4_zlmM>
* transformers: <https://www.youtube.com/watch?v=SZorAJ4I-sA>

Versuche, ob existierende KIs unser Problem lösen:

* ChatGPT, BingChat und Co: ohne Probleme (noch weiter ausführen)
* GPT4All:  
  - Mistral: klappt gut, auch auf DE (noch weiter ausführen)  
  - Falcon: Klappt kaum, lädt lange und stürzt ab, hängt sich oft auf (noch weiter ausführen)

Modell-Versuche:

* Orca 2:  
  - lädt erst lange (20Min) und ist dann abgebrochen  
  - im zweiten Versuch: Nach 72Min folgende Fehlermeldung:  
  ValueError: The current `device\_map` had weights offloaded to the disk. Please provide an `offload\_folder` for them. Alternatively, make sure you have `safetensors` installed if the model you are using offers the weights in this format.
* OpenChat: <https://huggingface.co/openchat/openchat_3.5>
* Mistral-7B-v0.1: <https://huggingface.co/mistralai/Mistral-7B-v0.1>
* zephyr-7b-beta: <https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>
* Llama-2-7b: <https://huggingface.co/meta-llama/Llama-2-7b>
* phi-1\_5: <https://huggingface.co/microsoft/phi-1_5>
* Llama-2-13b-chat-hf: <https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>
* GPT-2: <https://huggingface.co/gpt2>