**~~[Memory Complexity](https://arxiv.org/pdf/2305.19370.pdf) 읽고 추가 필요~~**

**~~3) Large Context Length를 위한 데이터 길이~~**

~~In-Context Instruct~~

~~학습 방법은 기본적으로 LoRA를 사용하되 Instruct learning 혹은 In-Context Instruct learning 중 골라서 한다. 둘 중 어느 것?~~

한국어 오픈 LLM 대부분이 polyglot-ko 모델을 사용했는 데, 해당 모델은 Korean hate speech 데이터셋 또한 사전학습시킨 모델로 예측 불가한 답변을 낼 수 있다.

XGen 써있지는 않지만 사용해 보니 한국어 데이터셋 또한 포함된 것으로 보인다. 사전학습 모델을 salesforce에서 instruction tune한 모델을 실험해 볼 수 있겠다.

[TODO] 오픈된 모델의 학습 방법, 데이터 등 비교 필요

XGen inst 모델의 성능이 좋아보인다. 하지만 한국어 데이터셋으로 pre-train된 내용이 많지 않아 한국어 finetuning으로는 적합하지 않을 수 있다. 확인이 필요하다.

<https://github.com/nlpai-lab/KULLM>

<https://github.com/melodysdreamj/KoVicuna>

<https://github.com/Beomi/KoAlpaca> - 4bit 양자화로 12.8B 모델을 Qlora로 사용한 학습하는 방법과 multi-gpu 학습 방법을 설명하는 코랩 링크를 포함하고 있다.

<https://github.com/kakaobrain/kogpt>

<https://huggingface.co/Salesforce/xgen-7b-8k-base>

<https://huggingface.co/EleutherAI/polyglot-ko-5.8b>

~~사전학습된 모델 선택 기준과 평가 방법은?~~

~~학습된 모델을 평가할 방법은?~~

ChatGPT에게 평가해 달라고 할 수 있다.

hallucination 문제는?

sensitive and lewd data가 출력된다면?

주제에 대한 범위를 정하고 벗어나는 질문은 못하게 막아놔야 할까?

싸게 학습데이터 만들기  
<https://github.com/22-hours/cabrita>  
[https://devocean.sk.com/blog/techBoardDetail.do?ID=164779&boardType=techBlog](https://devocean.sk.com/blog/techBoardDetail.do?ID=164779&amp;boardType=techBlog)  
PEFT - Parameter-Efficient Fine-Tuning: 모델의 일부 파라미터만 튜닝함으로써 모델의 성능을 적은 자원으로 높게 유지하는 방법론  
LoRA - PEFT기법 중 하나. 고정된 weight를 가진 pre-trained 모델에 학습이 가능한 rank decomposition 행렬을 모델 중간 중간 삽입한다.  
적은 양의 파라미터로 모델을 튜닝 가능하기 때문에 적은 수의 GPU로 빠르게 튜닝이 가능하다.  
FLAN - Finetuned LANguage Models are zero-shot learners. Instruction Tuning을 통해 학습에서 보지 못한 task도 할 수 있다.  
self-Instruct - LLM으로 데이터를 생성해서 그 데이터로 다시 LLM을 학습시키는 방법. Alpaca에서는 더 좋은 모델(GPT-3.5(text-davinci-003))등을 사용해 데이터를 생성했다. [<https://devocean.sk.com/blog/techBoardDetail.do?ID=164659>]

In-Context Instruct Learning - <https://arxiv.org/abs/2302.14691>