안녕하세요.

저희는 자연스럽고 윤리적인 일상 대화형 한국어 챗봇 모델 구현을 주제로 하고 있는 하계 LLM팀입니다. 저는 발표를 맡은 박성완입니다.

대부분의 LLM들은 대량의 말뭉치 데이터들을 한번에 학습하는 과정에서 혐오 표현이나 편향된 정보를 함께 학습할 수 있습니다. 또한 Opensource LLM들은 친근한 사람의 말투를 사용하지 않고, 기계적인 어투를 사용하여 어린아이들이 사용하기에는 무리가 있습니다.

저희는 이러한 잠재적 위험으로부터 안전한 LLM을 개발하고 남녀노소 누구나 친근하게 사용할 수 있는 LLM을 개발하는 것을 목표로하고 있습니다.

과제 목표는 다음과 같습니다.

먼저, 오픈소스 한국어 모델중에 가장 성능이 좋은 구름 모델을 baseline모델로 사용할 예정입니다. 두번째로는 Instruction tuning을 통해 사람 명령어의 이해도를 높이는 방식으로 혐오 표현 필터링을 진행할 예정입니다. 세번째로는 Reinforcement Learning from Human Feedback 방식을 통해서 인간이 판단한 적절성을 Reward로 활용하여 친근하고 정확한 답을 내놓는 방식으로 강화학습을 진행할 예정입니다. 마지막으로 Self-instruct 방식을 통해 GPT4로 데이터셋을 증강할 예정입니다.

Baseline 모델인 구름 모델에 대해서 소개하도록 하겠습니다. 구름 모델은 고려대학교 NLP & AI 연구실과 HIAI 연구소가 개발한 한국어 LLM입니다. 모델 파라미터 뿐만 아니라 데이터셋까지 전면 공개하였고 Polyglot-ko 기반 모델로 상업적 license를 보유한 모델입니다.

5.8B, 12.8B 기반 모델이 존재하고 구름 이전 한국어 SOTA 모델인 Ko-Alpaca의 성능을 놀라운 격차로 이겼다고 합니다.

다음으로 빠르고 효율적인 학습을 위한 방법론들을 소개하도록 하겠습니다.

먼저 LoRA 방식은 Pretrained model의 모든 weight를 finetuning하는 방법 대신, 모든 weight를 freeze하고 downstream task를 수행하기 위해 훈련 가능한 rank decomposition matrix를 추가 함으로써 훨씬 더 적은 리소스로 파인튜닝을 가능하게 합니다.

Self-Instruct 방식은 인간이 만든 적은 수지만 양질의 seed-task를 고성능 foundation model을 통해 비슷한 데이터들을 증강하는 방식을 의미합니다. 저희는 이러한 방법들을 사용하여 저희가 가지고 있는 적은 리소스의 제한을 해결할 예정입니다.

혐오 표현을 필터링하기 위해서 저희는 Instruction finetuning을 진행할 예정입니다. Instruction finetuning이란, pretrain되어 있는 고성능 모델에 적절한 <Instruction, output> 쌍을 넣어서 인간의 명령에 따르는 적합한 output이 나오도록 학습하는 방법을 의미합니다.

위와 같은 혐오 표현 데이터셋과 저희가 직접 제작한 혐오 표현에 해당하는 몇 개의 적절한 답변중 하나로 이루어진 pair를 instruction tuning 방식으로 구름 12.8B 모델에 학습시켜 혐오 표현이 들어왔을 때, 적절한 Output이 나오도록 할 예정입니다.

하지만 이러한 욕설 필터링과 다르게 모든 질문에 대한 자연스럽고 정확한 답변은 일관성 있는 몇 개의 답변으로 instruction tuning 처리가 힘듭니다. 따라서 저희는 인간이 판단한 적절성을 Reward로 활용하는 RLHF 방식으로 강화학습을 진행할 예정입니다.

RLHF 방식으로 friendly 하고 정확한 답변을 생성하는 과정을 설명하도록 하겠습니다. 먼저 friendly text instruction data를 구축합니다. 이 때 labeler가 데이터셋을 직접 검토하여 좋은 데이터만을 선별한 seed data를 GPT4로 증강합니다.

이렇게 증강된 최종 데이터셋으로 여러 한국어 모델들을 학습하여 Supervised Fine Tuning 모델을 제작할 예정입니다.

두번째로는 인간의 선호도를 반영한 Comparison 데이터 구축 및 Reward Model을 학습합니다. 먼저 만들어둔 SFT 모델들과 Chatgpt등의 모델에서 새로운 Prompt에 대응하는 4개 이상의 response를 생성합니다. Labeler는 직접 response의 선호도를 매겨 comparison data를 생성하고 이 데이터로 사람의 선호 답변을 생성하는 Reward Model을 학습합니다.

마지막으로 최종적으로 학습할 구름 12.8B 모델에 앞서 학습된 Reward Model을 강화학습의 reward funciton으로 사용하여 Proximal Policy Optimization 알고리즘으로 학습을 진행합니다.

최종적으로는 사용자의 입력에 사람의 기준에서 유연하고 정확한 답변을 생성하는 LLM을 제작할 수 있습니다.