이름: 조현수 (Hyunsoo Cho)
직위: 조교수 (Assistant Professor)
소속: 이화여자대학교 (Ewha Womans University)
기타소속:

강연제목: Towards Better Alignment in Large-Language Models: Current Research Trends (초거대 언어모델의 정렬 학습: 최신 연구 동향)

Abstract:

The ability of large language models (LLMs) to follow human instructions relies heavily on their alignment with human intentions. This talk reviews recent research trends on improving the alignment of LLMs with human goals. We explore key methods such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), highlighting recent studies that show these techniques can not only enhance alignment but also unlock new capabilities in LLMs. Finally, we discuss the remaining challenges in aligning LLMs more effectively with human intentions.

Brief Biosketch

(현) 이화여자대학교 인공지능학과 조교수
(현) 언어공학연구회 운영위원
(전) 서울대학교 컴퓨터 연구소 박사후 연구원
(전) 네이버 클라우드 방문연구원 (Foundation Research Team)
서울대학교 컴퓨터공학 박사
네이버 초거대 언어모델 HyperClova X 개발 참여