대형 언어모델의 심리 추론 능력 경량화를 통한 자가 인지치료 시스템 구현 연구

2025-1 AJOU SOFTCON

이 름 전환휘 지도교수 조현석

연구 배경 및 개발 목적

국내 정신건강 지표의 악화와 심리상담 서비스의 비대칭적 접근성 문제는, 누구나 일상적으로 사용할 수 있는 자가 심리 관리 도구의 개발 필요성을 부각시키고 있다. 인지행동치료(Cognitive Behavioral Therapy; CBT)는 정서 및 인지 왜곡 개선에 과학적 근거가 있는 심리치료법이지만, 일반인의 독립적 적용은 현실적으로 어렵다. 최근 대형 언어모델(LLM)의 발전은 CBT 지식 기반 질의응답 및 심리 추론 영역에서도 우수한 성능을 보여주고 있으며, 특히 CBT-BENCH 벤치마크를 통해 그 잠재력이 입증된 바 있다.
그러나 GPT-4o 및 Mistral-7B와 같은 고성능 LLM은 모바일 환경에서 실시간으로 활용하기에 연산 비용이 과도하게 높다는 한계를 지닌다. 본 연구는 이러한 한계를 극복하고자, LLM의 심리 추론 능력을 유지하면서도 경량화된 모델(Phi-2, Gemma-2b-it 등)로 distillation을 수행함으로써, 모바일 애플리케이션 수준에서 실행 가능한 심리 보조 시스템의 핵심 추론 엔진을 개발하는 것을 목적으로 한다.
아울러, 심리정보는 개인의 내면적 사고와 정서 상태를 반영하는 고위험 민감정보(Personal Mental State Data)로 간주되며, 이의 외부 전송 또는 상업 API를 통한 처리 과정에서 정보 유출 및 프라이버시 침해 가능성이 존재한다. 본 연구는 이러한 보안적 우려를 고려하여, distillation된 소형 모델(Student)이 클라이언트-사이드 또는 폐쇄형 로컬 서버 환경에서 직접 추론을 수행할 수 있도록 설계되었으며, 외부 LLM API 호출을 배제함으로써 정서정보 보호 및 사용자 익명성 보장을 실현한다. 이를 통해 단순한 경량화 모델 구현을 넘어, 실제 사용자 중심의 보안성과 독립성을 갖춘 심리 인터페이스 구축을 지향한다.
대형 언어모델(LLM; GPT-4o, Mistral-7B)은 고차 자연어 추론 능력을 보유하고 있으나, 계산 자원 측면에서 모바일 환경에 부적합하다. 이에 따라 본 연구는 teacher 모델의 reasoning trajectory를 소형 student 모델(Phi-2, Gemma-2b-it 등)에 내재화하기 위한 구조적 distillation 기법을 설계하고, 이를 CBT- BENCH Level II의 심리 추론 과제(CBT-PC)에 적용하여 심리 상담 보조용 경량 추론 엔진으로 확장 가능한지를 실험적으로 검증한다.

주요기술

본 연구는 인지행동치료(CBT)의 심리 추론 능력을 대형 언어모델(LLM)에서 소형 모델로 전이하기 위한 Explanation-Guided Distillation(EGD) 프레임워크 구현을 중심으로 한다. 단순 정답 복제가 아닌, LLM의 Chain-of-Thought(CoT) 기반 reasoning을 student 모델이 내재화하도록 설계하였다.
주요 기술 모듈은 다음과 같다.

1. CoT 기반 teacher reasoning 추출
GPT-4o 및 Mistral-v0.3-7B를 teacher로 설정하여 situation과 thoughts를 입력으로 받아, label과 함께 감정 반응, 인지 왜곡 해석, 핵심 신념 추론이 포함된 고차 reasoning을 생성한다. 이 데이터는 student 모델의 추론 모방 학습의 핵심 자원이 된다.

2. LoRA 기반 경량 학습
Phi-2 및 Gemma-2b-it에 대해 Hugging Face PEFT 라이브러리를 사용, 사전학습 파라미터는 고정하고 LoRA 모듈만 학습하여 전체 파라미터의 1% 이내로 reasoning 전이를 실현했다. 이는 모바일 환경에서의 효율성과 성능을 동시에 확보한다.

3. Reasoning Consistency Enforcement
동일 입력에 대해 student가 생성한 여러 reasoning 간 cosine 유사도를 측정, 낮은 경우 penalty를 부여하여 의미적 일관성과 논리 정합성을 강화한다. 이로써 reasoning 품질과 신뢰도를 향상시킨다.

4. Self-Refinement 루프
student가 생성한 reasoning을 다시 입력하여 스스로 검토·수정하는 2단계 추론 파이프라인을 구성하였다. 이 구조는 단순 모방을 넘어 reasoning 오류에 대 한 자기 피드백과 메타인지 능력을 실현한다.

5. FastAPI + Flutter 기반 UI 시스템
student 모델은 FastAPI 기반 추론 서버에 내장되며, Flutter 앱에서는 음성 입력(STT), 텍스트 출력, TTS 피드백을 통해 실시간 심리 인터페이스를 구현했다. 사용자는 발화를 통해 상황/사고를 입력하고, 모델은 reasoning과 label을 반환한다.

연구 및 개발내용

본 연구의 전체 구조는 LLM으로부터 생성된 심리적 추론 문장을 중심으로 한 다중 레이어 학습 파이프라인으로 설계되었다. 데이터 전처리, 출력 구조 정의, 학습 기법, 평가 메커니즘이 정밀하게 구성되었다.

1. CBT-PC 데이터셋 기반 Prompt-to-Distillation 구조 설계
   1. CBT-BENCH Level II의 Core Belief Classification 데이터를 기반으로 각 샘플에 대해 situation, thoughts, fine-grained belief 정보를 포함하는 고차 입력 포맷을 정의
   2. Teacher 모델에게는 “사고를 해석하고 그 배경 신념을 추론하시오”라는 형식의 semantically dense prompt를 제공하여, 단순 label이 아닌 reasoning을 유도
2. Teacher → Student 데이터 생성 및 distillation pipeline 구성
   1. reasoning이 포함된 출력을 기반으로 input-output pair를 생성하여 student 모델 학습용 corpus 구성
   2. reasoning 문장은 감정 표현, 인지 왜곡 해석, 신념 추론이라는 3단 논리 구조를 갖도록 일관되게 설계되었으며, 모델이 정답뿐 아니라 사고 과정 자체를 학습하도록 유도
3. Multi-level Supervision 및 Loss Structuring
   1. Cross-Entropy loss에 더해 reasoning간 coherence loss와 label- reasoning alignment loss를 추가하여, 단순 정확도 이상으로 reasoning validity를 반영한 학습 구조 구현
   2. 데이터 불균형 완화를 위해 focal loss 및 label frequency normalization을 병행 적용
4. 시스템 구현
   1. 학습된 모델은 FastAPI 기반 추론 서버로 배포되며, Flutter로 구현된 프론트엔드와 음성 입력 기반 인터페이스로 연결됨
   2. 사용자의 실시간 발화를 text로 변환한 후 상황/사고 포맷으로 재구성하고, 모델은 reasoning + label 응답을 생성함

결과 및 분석

실험은 reasoning 능력의 전이 가능성과 출력의 신뢰도를 평가하기 위해 정량적 지표뿐 아니라 언어적 일관성, 심리적 해석력, reasoning path의 질적 구조까지 포함하는 총체적 분석 프레임워크를 적용하였다.

Quantitative Analysis (Multi-label Classification Metrics)
- 실험 조건 : Instruction Only / +CoT Prompt / +Consistency / +Self- Refinement
- 주요 지표 : Macro F1, Reasoning BERTScore, GPT-4 기반 인적 평가

조건	Macro F1	BERTScore	GPT-4 평가 (5점 만점)	일관성 지수
Instruction Tuning Only	0.57	-	-	-
+CoT Reasoning	0.63	0.74	4.1	0.61
+Consistency Loss	0.66	0.81	4.6	0.78
+Self-Refinement	**0.68**	**0.84**	**4.8**	**0.82**

Qualitative Analysis
- CoT reasoning이 없는 모델은 모든 label을 반복 출력하는 경향을 보였으며, GPT-4 기반 평가는 평균 3.1점에 그침
- CoT 포함 후에는 자동사고에 근거한 감정 해석이 포함되었으며, reasoning 문장이 상황과 label 간 인과성을 반영
- Self-Refinement 후 reasoning 구조의 내적 논리 일관성과 감정-신념 추론 간 연결성이 증가됨
- 출력 reasoning은 LIME 기반 토큰 기여도 분석 결과, 상황과 감정 표현에 대한 주의(attention) 집중도가 1.3배 향상됨