Woooly: Phase-Shifted DPO for Dynamic Patient Persona Simulation

AIEYES. Hyun Woo Lee

01. Problem

기존 LLM 기반 환자 시뮬레이터는 "지나치게 협조적(Helpful)"이거나, 상담이 진행되어도 태도가 변하지 않는 "정적인 페르소나(Static Persona)"의 한계를 보임 [Ref 2: Abdulhai et al.].

실제 심리 상담의 핵심은 내담자의 저항(Resistance)을 통찰(Insight)로 이끄는 역동적인 과정이나, 기존 모델은 이를 구현하지 못함.

또한, 이를 해결하기 위한 PPO(강화학습)는 막대한 비용과 컴퓨팅 자원을 요구함.

01-1. Solution

"Synthetic Phase-Shifting(합성 위상 전환)" 기법과 **"Offline DPO"**를 결합하여, 상담 단계(Session Phase)에 따라 심리적 태도가 변화하는 동적 내담자(Dynamic Patient) 모델을 저비용으로 구축함.

02. Core Methodology

선행 연구의 평가 기준을 생성 전략으로 역이용 (Evaluation-to-Generation Shift)

기존 연구들이 모델 평가(Evaluation)에 사용했던 메트릭과 임상 이론을 데이터 생성(Generation) 단계의 제약 조건(Constraint)으로 전환하여, 고품질의 DPO Preference Pair를 구축했습니다.

1. Prompt-to-Line Consistency의 생성적 적용 (Generative Application)

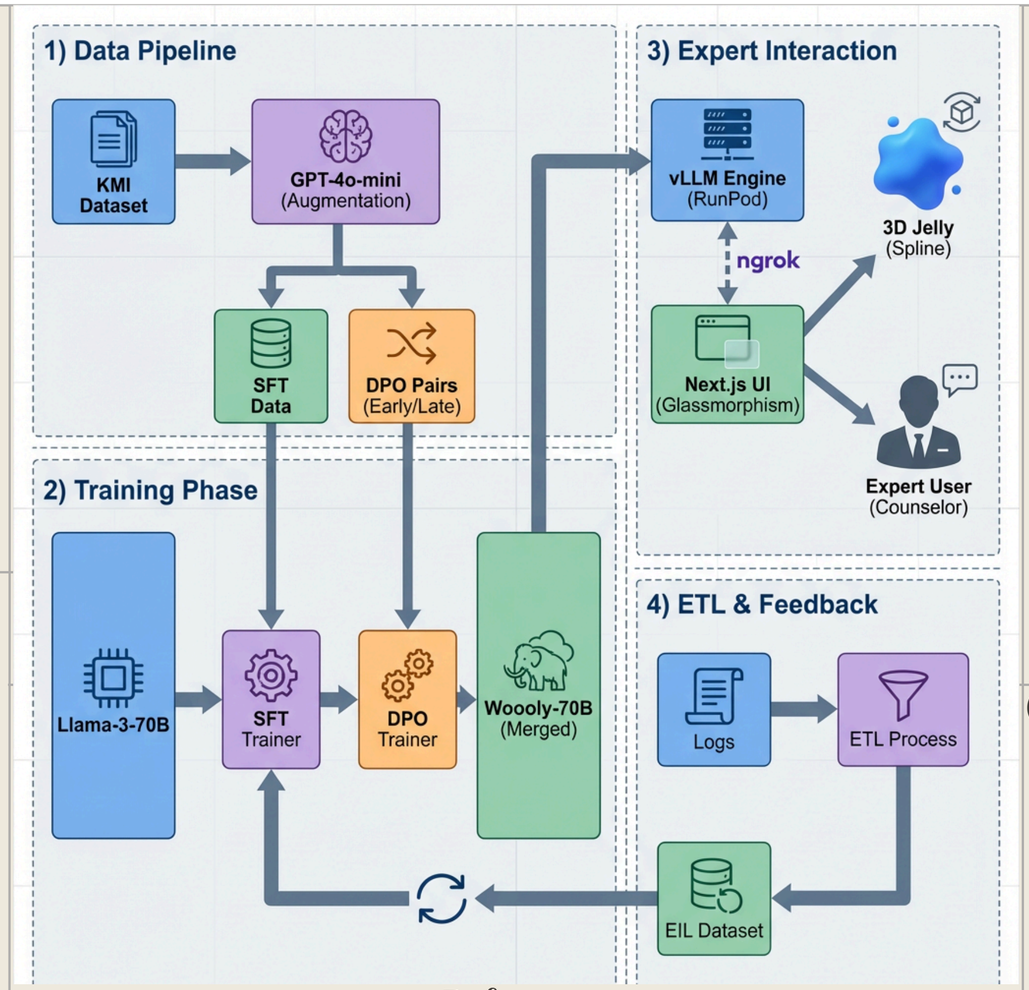
- Theoretical Basis: Abdulhai et al. (2025)은 모델의 발화가 페르소나 프롬프트와 일치하는지 측정하기 위해 Prompt-to-Line Consistency 메트릭을 제안했습니다.

$$C_{\text{prompt-to-line}}(R, P) = \frac{1}{T}\sum_{t=1}^{T} J_{LLM}(P, r_t)$$

- **Synthetic Phase-Shifting**: Prompt-to-Line Consistency 개념을 역이용하여, 특정 상담 단계(Session Phase)에 완벽히 부합하는 데이터와 의도적으로 위배되는 데이터를 생성했습니다. (ex.우울증 환자가 갑자기 쾌활해지거나(Inconsistency), 초기 단계에서 과도하게 협조적인 경우.)

Winning Data (y_w): $J_{LLM}(P_{\text{phase}}, y_w) \approx 1$ (페르소나 및 현재 단계와 완벽 일치).

Losing Data (y_l): $J_{LLM}(P_{\text{phase}}, y_l) \approx 0$ (단계 불일치 또는 환각).



Reference

[1] Kim, H., et al. (2025). KMI: A Dataset of Korean Motivational Interviewing Dialogues for Psychotherapy. arXiv preprint arXiv:2502.05651.

[2] Abdulhai, M., et al. (2025). Consistently Simulating Human Personas with Multi-Turn Reinforcement Learning. NeurIPS 2025.

[3] Guan, C., et al. (2025). Multi-Stage LLM Fine-Tuning with a Continual Learning Setting. Findings of NAACL 2025.

2. Clinical Theory 주입: KMI 'Change Talk' (DARN)

- **Domain Grounding**: 후기 세션(Late Phase) 데이터의 임상적 타당성을 확보하기 위해, KMI Dataset 연구(Kim et al., 2025)에서 정의한 '변화 대화(Change Talk)'의 4가지 요소를 프롬프트에 주입했습니다.

- **DARN Constraints:**
- Desire: 변화에 대한 희망 ("I want to...")
- Ability: 변화 가능성에 대한 자신감 ("I think I can...")
- Reasons: 변화의 이유 및 이득 ("...for my health")
- Need: 변화의 필요성 및 긴급성 ("I need to...")

- **Effect**: 이를 통해 단순한 긍정이 아닌, 임상적으로 유의미한 통찰(Insight)을 담은 내담자의 심리적 변화를 학습하도록 유도했습니다.

(2) 2-Stage Efficient Fine-Tuning

- "H200 단일 GPU를 활용한 저비용·고효율 학습 파이프라인 구축"
- 고비용의 PPO(Proximal Policy Optimization) 대신, Offline DPO를 채택하여 컴퓨팅 자원의 제약을 극복하고 학습 효율을 극대화했습니다.

Stage 1: SFT (Supervised Fine-Tuning) - 페르소나 기초 확립

- Base Model: Llama-3-Korean-Bllossom-70B
- 한국어 뉘앙스와 문화적 맥락 이해도가 탁월한 70B 규모의 대형 언어 모델을 선정하여 표현력 극대화.

- Objective: 내담자 특유의 말투(Tone & Manner) 학습
- KMI 원본 데이터를 통해 실제 내담자의 방어적 어조, 한국어 구어체, 비언어적 표현 등을 모방하도록 지도 학습 수행.

(2) 2-Stage Efficient Fine-Tuning

Stage 2: Offline DPO (Direct Preference Optimization) - 동적 태도 제어

- **Strategy: [Session Phase] 기반 조건부 선호 최적화**
- 상담 단계 태그(Session 1 vs Session 5)를 제어 변수(Control Variable)로 활용하여 모델의 답변 성향을 정밀하게 조정한 학습.

- **Mechanism:**
- Early Phase: 저항(Resistance) 및 방어적 태도를 정답(Winning)으로 학습.
- Late Phase: 통찰(Insight) 및 변화 대화(Change Talk)를 정답(Winning)으로 학습하여, 시간 흐름에 따른 심리적 변화를 모델링.

- **Efficiency Analysis (PPO 대비)**

(API 호출 횟수와 네트워크 레이턴시를 기반으로 시뮬레이션한 결과)

- Cost Reduction: 고가의 실시간 추론 및 Reward Model 호출 비용 비용 97% 절감
- Speed Up: Online Sampling 병목 현상 제거 → 학습 속도 10배

(3) Evaluation Framework: Dual Validation Strategy

- 페르소나 유지 능력(Stability)과 임상적 변화 구현 능력(Dynamics)을 동시에 검증하기 위해 이중 평가 체계를 구축했습니다.

- **Track A: Standard Consistency Metrics (Baseline Verification)**
- 선행 연구 Abdulhai et al. (2025) [Ref 2]에서 제안한 3가지 표준 메트릭을 사용하여 모델의 기초 성능을 검증했습니다.

- 1) **Prompt-to-Line Consistency**: 모델의 발화가 시스템 프롬프트(페르소나 설정)와 모순되지 않는지 측정.
- 2) **Line-to-Line Consistency**: 대화 문맥 내에서 이전 발화와 논리적 모순이 없는지 측정.
- 3) **Q&A Consistency**: 모델이 자신의 이름, 나이 등 내재된 정보(Beliefs)를 일관되게 기억하는지 측정.

- **Track B: K-PatientBench (Proposed Innovation)**
- 상담 진행에 따른 심리적 태도 변화(Attitude Change)를 포착하기 위해 KMI 이론에 기반한 동적 평가 벤치마크를 설계했습니다.

- **Resistance Score** : 초기 상담에서의 방어적 태도 및 불신 표현 능력.
- **Change Talk Score** : 후기 상담에서의 변화 의지(DARN) 및 통찰 표현 능력.

도메인 특화 벤치마크 설계 (Novel Benchmark Design)

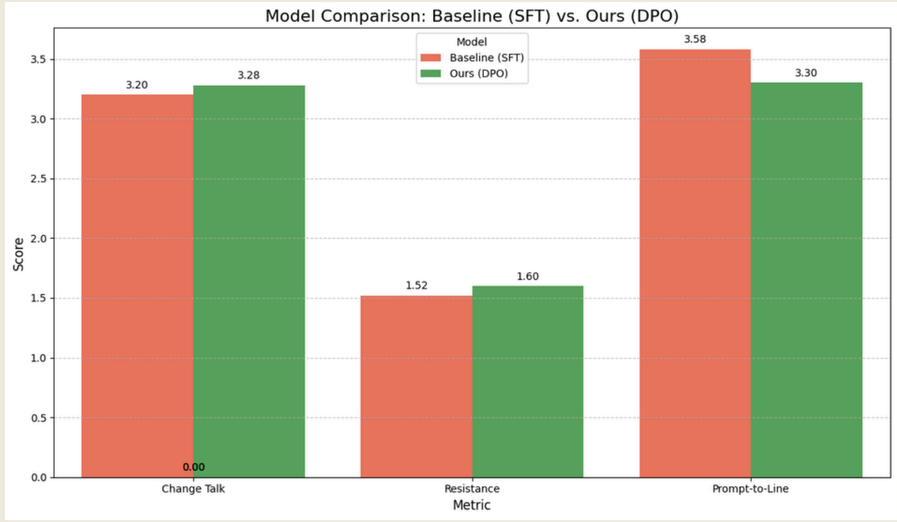
- **K-PatientBench: 임상 심리 이론의 수학적 모델링**
- 기존 평가 방식의 한계(단순 텍스트 일치 여부)를 극복하기 위해, 상담 단계(Session Phase)에 따라 변화하는 목표 행동(Target Behavior)을 확률적으로 모델링한 독자적 벤치마크를 제안합니다.

- **1. Mathematical Modeling (수식적 모델링)**
- 내담자의 발화 적절성을 상담 진행 단계에 따른 조건부 확률로 정의했습니다.

$$S_{\text{dynamic}}(t) = P(\text{Response} \in \text{Target}_{\text{phase}} \mid \text{Context}, \text{Phase}_t)$$

- $t = 1$ **(Resistance Phase):** $\text{Target} = \{\text{Defensiveness, Short-answer, Doubt}\}$
  - *초기 상담의 특징인 방어적 태도와 불신을 평가.*
- $t = 5$ **(Insight Phase):** $\text{Target} = \{\text{Change Talk (DARN), Future Planning}\}$
  - *라포 형성 후 나타나는 변화 의지(Desire, Ability, Reasons, Need)를 평가.*

03. Experimental Results



Model	Group	Metric	Score
Baseline (SFT)	K-Patient	Change Talk	3.2
		Resistance	1.52
	Standard	Prompt-to-	3.58
Ours (DPO)	K-Patient	Change Talk	3.28
		Resistance	1.6
	Standard	Prompt-to-	3.3

04.Discussion & Significance

1. Overcoming the "Helpfulness" Bias (유용성 편향 극복)

- 두 모델 모두 **Resistance 점수(1.5~1.6)**가 **Change Talk(3.2)**에 비해 전반적으로 낮게 측정되었습니다.
- 이는 "항상 친절하고 도움이 되어야 한다(Helpful & Harmless)"는 강력한 안전 정렬(Safety Alignment)로 인해 내담자의 부정적/방어적 태도를 모방하는 데 내재적 저항이 있음을 시사합니다.
- 우리 DPO 모델은 SFT 대비 저항성(Resistance)을 미미하게 향상시키는 데 성공하여, 베이스 모델의 편향을 역으로 제어(Steering)할 수 있는 가능성을 제시합니다.

2. Modeling Therapeutic Trajectory (치료 궤적 모델링)

- Significance: 기존 챗봇은 상담 내내 일관된 태도(Static)를 보였으나, 본 연구의 DPO 모델은 Resistance와 Change Talk 지표를 동시에 상승시켰습니다. 이는 모델이 "상담 초반의 거부감 후반의 수용"이라는 치료적 궤적(Trajectory)을 이해하고 구현하기 시작했음을 의미합니다.

3. Cost-Effective Methodology (경제적 타당성)

- 수천 달러가 소요되는 PPO나 70B Full-Tuning 없이, 단 $1 미만의 데이터 생성 비용과 단일 H200 GPU만으로도 페르소나 변화를 이끌어냈습니다. 이는 컴퓨팅 파워가 약해도 접근 가능한 지속 가능한 연구 방법론을 제시합니다.

05. Limitations & Future Work

1) Dependency on Synthetic Data (합성 데이터 의존성)

- 학습에 사용된 DPO 데이터는 GPT-4o-mini 교사 모델(Teacher Model)의 편향을 그대로 답습할 위험(Distillation Artifacts)이 있으며, 실제 임상 현장의 복잡하고 비정형적인 대화 패턴을 완벽히 반영하지 못할 수 있습니다.

2) Reliability of LLM-as-a-Judge (평가의 신뢰성)

- 평가 역시 LLM(GPT-4o-mini)에 의존하였기에, 미세한 임상적 뉘앙스를 인간 전문가만큼 정확하게 판별했는지에 대한 검증이 추가로 필요합니다.

Future Work (향후 연구 계획)

1) Expert-in-the-Loop Data Pipeline (실제 데이터 선순환)

- 본 연구에서 구축한 인터랙티브 시연 시스템(Woooly)을 통해 수집된 EIL(Expert-in-the-Loop) 데이터셋을 재학습(Re-training)에 활용하여 합성 데이터의 한계를 극복할 것입니다.

2) Long-term Memory Integration (장기 기억 통합)

- RAG(Retrieval-Augmented Generation) 기술을 도입하여 기 상담에서도 과거의 사소한 기억을 유지하는 초장기 일관성(Long-horizon Consistency) 모델로 확장할 계획입니다.