데이터 유도형 공감 페르소나 벡터 추출 기법

정인재[1], 김태훈[1], 박진영[2], 박천음[1]*

[1]국립한밭대학교, [2]성균관대학교

jeongij@edu.hanbat.ac.kr, thkim@hanbat.ac.kr, jy.bak@skku.edu, *parkce@hanbat.ac.kr

Data-Driven Empathy Persona Vector Extraction Technique

Injae Jeong[1], Taehoon Kim[1], Jinyoung Park[2], Cheoneum Park[1]*

[1]Hanbat National Univ., [2]Sungkyunkwan Univ.

요 약

최근 LLM은 RLHF로 인해 방어적이고 피상적인 공감에 머무르는 한계가 있다. 본 연구는 이를 해결하고자 ED-S 데이터셋과 다층 필터링 기반의 데이터 유도형 공감 페르소나 벡터 추출 기법을 제안한다. 이는 Few-shot과 3중 필터링으로 대조군 생성을 엄격히 통제하여, 인간 고유의 공감 패턴을 벡터에 정밀하게 투영한다. 실험 결과에 따르면 제안 기법은 모델의 고유한 사실성과 유창성을 훼손하지 않으면서도 실제 인간의 공감 양상이 반영된 응답으로 유도됨을 입증한다.

I. 서론

최근 거대언어모델(Large Language Model, LLM)이 대화형 인공지능의 주요 기반 모델로 자리 잡았으나, RLHF (Reinforcement Learning from Human Feedback) 과정으로 인해 상투적인 위로나 방어적인 해결책 제시에 치중하는 경향을 보인다[1, 2]. 이러한 경향은 사용자와 AI 간의 깊이 있는 라포(Rapport) 형성을 저해한다. 이에 최근에는 모델 가중치를 동결한 채 추론 과정에서 활성화 값만을 조정하여 특정 행동을 유도하는 페르소나 벡터(Persona Vector) 방법론이 제안되었다[3]. 그러나 페르소나 벡터 추출을 위해 LLM이 생성한 합성 데이터에 전적으로 의존한다는 한계가 있다. 공감과 같은 고차원적 감정 영역에서 합성 데이터는 앞서 지적한 모델의 기계적 편향을 되풀이할 뿐, 실제 인간의 복합적인 감정 기제를 온전히 반영하지 못한다. 이에 본 논문은 실제 인간 공감 대화 데이터셋인 Empathetic Dialogues (ED) [4]를 활용한 데이터 유도형 공감 페르소나 벡터 추출 기법을 제안한다.

II. 제안 방법

본 논문에서는 기존의 합성 데이터 의존도를 낮추고 신뢰성을 높이기 위해, ED 데이터셋으로부터 ED-S(Empathetic Dialogues for Steering)를 재구성한다. 이를 기반으로 필터 과정을 거친 공감 페르소나 벡터 추출 기법을 제안한다.

II.I. ED-S 데이터셋 구축

ED-S 데이터셋은 대조군의 활성화 차이를 포착하는 데 최적화된 사용자 입력-공감 응답 쌍의 단일 턴 구조를 이루며, 특정 감정에 대한 편향을 방지하고자 32가지 감정 레이블이 균등하게 분포하도록 구성한다. 이후 데이터의 신뢰성을 확보하기 위해 인간 검수(human evaluation)를 수행한다. 이 과정에서 문맥이 불분명하거나 공감 표현이 모호한 샘플은 배제하고, 확실한 공감을 포함한 고품질 데이터만을 선별한다. 구축된 데이터셋은 용도에 따라 페르소나 벡터 추출을 위한 Extraction Set, 최적 레이어 선별을 위한 Valid Set, 그리고 모델의 일반화 성능 평가를 위한 Test Set으로 구성되며, 통계는 [표 1]과 같다.

데이터셋 구분	샘플 수	설명
Extraction Set	3,000	감정 레이블별 ≈ 94개
Valid Set	64	감정 레이블별 2개
Test Set	244	감정 레이블별 ≈ 8개
합계	**3,308**	32개 감정 레이블 균등 분포

표 1: ED-S 데이터셋의 통계

II.II. 공감 페르소나 벡터 추출 및 검증

본 논문은 기존 페르소나 벡터 추출 기법[3]을 기반으로 하되, 데이터셋 특성을 반영한 프롬프트와 검증 파이프라인을 구축하여 추출되는 벡터의 품질을 개선한다.

대조적 응답 생성 및 다층 필터링. 추출용 데이터셋으로부터 대조적 활성화를 유도하기 위해 긍정 응답 R_{pos}과 부정 응답 R_{neg}을 생성한다. 모델이 ED-S 데이터셋의 실제 인간의 공감 특징을 효과적으로 모방하도록 페르소나(persona)를 설계하고, 동일 레이블의 타 샘플을 예시로 제공하는 few-shot 프롬프팅을 적용하여 R_{pos}를 생성한다. R_{neg}는 동일한 문맥에서 감정을 배제하고 사실만을 전달하도록 지시하여 활성화 차이를 만든다. 생성된 R_{pos}의 신뢰성을 확보하기 위해, gold response를 기준으로 최소 길이 보장, 스타일 유사도(ROUGE-L) 측정, 의미적 유사도(BERTScore) 검증, 그리고 GPT-4o-mini를 활용한 LLM-as-a-Judge[5]로 구성된 검증 파이프라인을 구축한다. 이러한 다층적 필터를 통과하여 실제 벡터 산출에 사용되는 대조군의 집합을 $\mathcal{D}_{valid}$로 정의한다.

*Corresponding author

실험 방법 (Method)	강도 (α)	의미적 유사도		어휘적 유사도	정성적 평가 (LLM Judge)		
		BERT	S-BERT	ROUGE-L	공감성	사실성	유창성
Baseline	-	0.850	0.387	0.108	4.41	4.23	4.69
Baseline (Persona)	-	0.859	0.384	0.138	4.48	4.22	4.78
LoRA Fine-tuning	-	0.871	0.362	0.192	2.64	3.56	3.86
Previous Work	1.0	0.840	0.379	0.090	4.74	4.04	4.57
	1.5	0.835	0.349	0.083	**4.75**	3.89	4.37
Ours	1.0	0.854	**0.398**	0.125	**4.66**	**4.20**	**4.86**
	1.5	0.855	0.391	0.127	4.58	4.16	4.84
Ours w/o Few-shot	1.0	0.856	0.399	0.126	4.41	4.29	4.81
	1.5	0.858	0.400	0.135	4.37	4.22	4.78

표 2: LLaMA-3-8B-Instruct 실험 결과.

레이어별 페르소나 벡터 후보 도출 및 레이어 선별. 검증된 샘플 집합 $\mathcal{D}_{valid}$를 바탕으로 각 레이어별 페르소나 벡터를 생성한다. 레이어 l에 대한 페르소나 벡터 후보 $\vec{v}^{(l)}$은 식 1과 같이 전체 샘플의 응답 생성 시 발생하는 내부 활성화 값의 차이를 산술 평균하여 계산한다.

$$\vec{v}^{(l)} = \frac{1}{|\mathcal{D}_{valid}|} \sum_{i \in \mathcal{D}_{valid}} \left(\overline{\mathbf{h}}_{pos,i}^{(l)} - \overline{\mathbf{h}}_{neg,i}^{(l)} \right) \quad (1)$$

이후, 최적의 개입 레이어 선별을 위해 후보 벡터를 각 레이어에 순차적으로 적용하여 응답을 생성한다. 레이어별 생성 응답과 gold response 간의 어휘적, 의미적 유사도 및 LLM-as-a-Judge 점수를 합산하여 복합 점수를 산출한다. 이 중 가장 높은 점수를 기록한 레이어를 최종 페르소나 벡터 레이어로 선정한다.

III. 실험

본 논문은 LLaMA-3-8B-Instruct를 백본으로 사용하고 ED-S Test Set을 통해 성능을 검증한다. 모든 실험에서 모델의 입력은 상황과 사용자 발화로 구성되며, 이에 대해 생성된 응답을 평가한다. 비교 실험군은 (1) Baseline 및 Baseline (Persona): 벡터 적용 없이 세부 페르소나 부여/미부여 설정, (2) LoRA Fine-tuning: ED-S 로 LoRA 튜닝한 지도 학습 방식, (3) Previous Work: 기존 페르소나 벡터 추출 기법[3]으로 추출한 벡터 적용, (4) Ours: 본 논문에서 제안한 Few-shot 및 3중 필터링 기반으로 추출한 벡터 적용, (5) Ours w/o Few-shot: Ours의 벡터 추출 과정에서 Few-shot 제거 방식(ablation study)으로 구성된다. Ours의 필터 임계값은 최소 길이 80%, ROUGE-L 0.125, BERTScore 0.855로 설정하는데, 이는 임계값 상향 시 급격하게 늘어나는 추출 시간과 벡터의 정제 수준 사이의 구조적 Trade-off를 고려한 수치이다. 해당 임계값은 고성능 LLM(GPT-5.1, Gemini-3-pro)이 gold response의 의미와 스타일을 모방할 때 보이는 유사도 점수를 참고하여 산정한다.

평가는 어휘 일치도를 보는 ROUGE-L과 의미 유사도를 측정하는 BERTScore, S-BERT를 사용한다. LLM-judge를 위해 GPT-4o-mini를 활용하여 0-5점 범위로 평가를 수행하며,평가 항목으로 감정 및 상황 반영도를 보는 공감성, 환각 여부를 판단하는 사실성, 어조의 자연스러움을 측정하는 유창성을 정의한다.

III.I. 실험 결과 및 분석

[표 2]는 제안 방법론(Ours)이 공감 성능을 유지하면서도 모델의 기존 능력을 안정적으로 보존함을 보인다. 구체적으로 Previous Work ($\alpha = 1.5$)는 가장 높은 공감성(4.75)을 보이나, 사실성(3.89)과 유창성(4.37), 의미적 유사도가 Baseline(Persona)에 비해 감소한다. 이는 선행 연구[3]에서 추론 단계 스티어링(Inference-time steering)의 한계로 명시된 모델 성능 저하 현상으로 보인다.

반면, Ours ($\alpha = 1.0$)는 공감성에서 4.66, 사실성에서 4.20을 유지하면서 유창성에서 4.86으로 가장 좋은 성능을 달성한다. 여기서 공감성 성능은 Baseline (Persona)의 4.48을 넘어서고 Previous Work(4.75)에 근접한다. 또한, S-BERT 점수는 다른 모델에 비해 높은 0.398로, 문맥의 의미를 훼손하지 않으면서 실제 인간같은 공감 어조 생성이 가능하다. 이에 따라, 제안 방법은 공감 영역에서 별도의 추가 학습과 복잡한 세부 페르소나 지시 없이 Steering Vector만으로 모델을 효율적으로 제어하며, 고품질의 응답 생성이 가능함을 알 수 있다.

Ablation Study에서 Few-shot 적용 시 공감성이 대폭 향상된 반면 사실성이 소폭 감소(4.29→4.20)한 점은 사실적 엄밀성보다 정서적 단서와 관계적 유대에 더 민감해지는 인간 공감 기제의 특성이 벡터 공간에 투영되어 나타난 구조적 Trade-off로 해석된다[6]. 한편, LoRA Fine-tuning는 소규모 데이터 과적합으로 인해 공감성과 사실성이 모두 하락하여, 데이터가 부족한 상황에서는 제안 기법과 같은 추론 단계의 제어가 더 효과적일 수 있음을 보인다.

IV. 결론

본 논문은 합성 데이터에 의존한 기존 방식을 벗어나, 실제 인간의 공감을 투영하는 데이터 유도형 공감 페르소나 벡터 추출 기법을 제안한다. 실험 결과, 제안 방법은 모델의 사실성을 훼손하지 않으면서도 자연스러운 공감 어조를 구현하여, 실제 데이터 기반 접근의 우수성을 입증한다. 향후에는 인간 평가를 도입하고, 한국어 및 다문화 환경으로 연구를 확장하여 방법론의 범용성을 확보할 계획이다.

참 고 문 헌

[1] A. D. Lindström, L. Methnani, and L. Krause, "Helpful, harmless, honest? sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback," *Ethics and Information Technology*, vol. 27, p. 28, jun 2025.

[2] Y. Chen, X. Xing, J. Lin, H. Zheng, Z. Wang, Q. Liu, and X. Xu, "SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, (Singapore), pp. 1170–1183, Association for Computational Linguistics, 2023.

[3] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey, "Persona vectors: Monitoring and controlling character traits in language models," *arXiv preprint arXiv:2507.21509*, sep 2025.

[4] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 5370–5381, Association for Computational Linguistics, July 2019.

[5] Y. Liu *et al.*, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of EMNLP 2023*, pp. 2511–2522, 2023.

[6] A. J. Martingano and S. Konrath, "How cognitive and emotional empathy relate to rational thinking: empirical evidence and meta-analysis," *The Journal of Social Psychology*, vol. 162, no. 1, pp. 143–160, 2022.