한국어-영어 공감대화 데이터셋과 성격을 기반으로 한 언어모델 평가

이영준[1] 현종환[1] 이도경[2] 성주원[2] 최호진[1]
[1]한국과학기술원 전산학부 [2]KT
{yj2961,hyeon0145,hojinc}@kaist.ac.kr {dokyong.lee,jwsung}@kt.com

Language Model Evaluation Based on Korean-English Empathetic Dialogue Datasets and Personality

Young-Jun Lee[1] JongHwan Hyeon[1] DoKyong Lee[2] Joo-Won Sung[2] Ho-Jin Choi[1]
[1]School of Computing, KAIST [2]KT Corporation

요 약

본 연구는 다양한 대규모 언어 모델들의 한국어/영어 공감 대화 생성에서 성능을 실험적으로 비교 분석하는 것과 개인의 성향과 공감 사이에서의 상관 관계를 실험적으로 분석하는 것을 목표로 한다. 이를 위해, 한국어 공감 대화 데이터셋인 KorEmpatheticDialogues 를 구축하였고, personality-aware prompting 방법을 제안한다. 실험을 통해, 총 18개의 언어 모델들 간의 공감 대화 생성 성능을 비교 분석하였고, 개인의 성향에 맞춤형 제공하는 공감이 더 상호작용을 이끌어낼 수 있다는 점을 보여준다. 코드와 데이터셋은 게재가 허용되면 공개할 예정이다.

주제어: 대규모 언어 모델, 공감 대화 생성, personality, MBTI

1. 서론

일상생활 대화, 상담 대화 도중에 내담자의 정서적 상황을 이해하여 공감을 하는 것은 내담자와의 깊은 관계 형성에 있어서 중요하다. 마찬가지로, 인공지능 어시스턴트가 사회적이고 사용자와 상호작용을 효과적으로 하기 위해서는, 인공지능 어시스턴트가 공감을 하거나 내담자의 입장에서 상황을 바라보는 등의 사회적 추론 능력 (social reasoning capability) 을 함유하고 있어야 한다 [1]. 그렇다 보니, 2018년에 EmpatheticDialogues 벤치마크 데이터셋 [2]이 제안됨에 따라, 많은 연구들이 대화 시스템이 사용자의 정서적 상황에 공감을 할 수 있도록 다양한 시도들이 이루어지고 있다 [3, 4, 5, 6, 7, 8]. 그러나, 기존 연구들은 "공감을 잘하자" 에만 초점을 맞추고 있고 개인의 성향 정보를 고려하고 있지 않는다. 예를 들어, 감정 지수가 높은 내담자의 경우에는 감정적인 리액션을 더 우선으로 해주는 사람의 응답이 더 공감을 해주고 있다고 느낄 수 있을 것이다. 즉, 본 연구에서는 개인의 성향에 따른 맞춤형 공감을 해야한다고 생각되며, 이를 위해 personality 중 하나인 MBTI [9]가 주입된 프롬프팅을 통해 personality 와 empathy 간의 상관 관계를 실험적으로 분석하고자 한다.

최근에, instruction tuning 혹은 reinforcement learning from human feedback (RLHF) 등의 기술로 학습된 대규모 언어 모델들 (예: InstructGPT [10], ChatGPT [11] 등)이 다양한 태스크에서 괄목할만한 성능을 보이고 있다 [12, 13]. 최근에는, social knowledge 태스크인, social bias 와 toxicity [14], Theory-of-Mind (ToM) [15], 공감 대화 [7], 에서 대규모 언어 모델의 성능 변화를 실험적으로 보여주고 있다. 본 연구의 주된 태스크인 공감 대화에서는, 기존 연구에서 GPT-3 [16] 모델에 대해서 공감 대화에서의 성능을 다양한 퓨샷 셋팅과 다양한 정량적 지표를 통해 평가하였다. 그러나, 기존 연구는 다양한 언어 모델들에 대해서 holistic 한 비교 분석을 진행하지 않았다. 또한, 현존하는 공감 대화 데이터셋들이 영어로 이루어져 있어, 한국어에서 대화 시스템의 공감 대화 능력을 평가하거나 분석하기는 어렵다.

본 연구에서는 다양한 대규모 언어 모델들의 한국어/영어 공감 대화 생성에서의 성능을 분석하는 것과 personality 와 공감 사이의 상관관계를 실험적으로 분석하는 것을 목표로 한다. 이를 위해, 번역기를 통해 한국어 공감 대화 데이터셋인 KorEmpatheticDialogues 를 자동 구축한다. 또한, 심플하며 직관적인 personality-aware prompting 방법을 제시하며, MBTI 타입 별로 대표 문장을 구축하여 이를 프롬팅을 통해 언어 모델에게 성향을 주입하고자 한다. 그리고 GPT-4 기반의 자동 평가 방법을 통해 대규모 언어 모델들의 공감 대화 성능을

표 1. KorEmpatheticDialogues 예시

Emotion: Excited
Situation: 크리스마스 콘서트 티켓을 받았습니다.
Dialogue:
발화자 1: 콘서트에 빨리 가고 싶어요.
발화자 2: 어떤 콘서트요?
발화자 1: U2 콘서트요. 티켓이 정말 비싸서 갈 수 없을 거라고 생각했는데 어떻게든 갈 수 있었어요!!!
발화자 2: 와우, 멋지네요! 실제 콘서트에 가본 적이 없어요.

표 2. KorEmpatheticDialogues 데이터셋 통계

Type	# Dialog	# Utter	Avg. Utter. Len	Avg. Utter/Dialog
train	19,531	80,508	13.45	4.12
valid	2,769	11,476	14.50	4.14
test	2,547	10,518	15.33	4.13
total	24,847	102,502	14.43	4.13

holistic 하게 측정한다. 포괄적인 실험 분석을 통해, 오픈 소스 언어 모델보다 공개되지 않은 모델들인 ChatGPT 와 GPT-4 가 전반적으로 성능이 높은 것을 확인하였다. 더 나아가, MBTI 타입별 공감과의 상관관계 분석을 통해, F, P, I, N 타입이 높은 공감을 이끌어내는 것을 알 수 있었다.

본 연구의 기여하는 점은 다음과 같다. (1) 한국어 공감 대화 데이터셋(KorEmpatheticDialogues)을 구축함에 따라 한국어에서 공감 대화 연구의 촉진을 이끌어 낼 수 있으며, (2) 한국어/영어 대규모 언어 모델의 공감 대화 생성 능력을 심도 있게 분석함에 따라 공감 대화의 능력을 위해 어떤 점을 디자인하는 게 좋을지에 대한 부분도 고찰시킬 수 있고, (3) 성향을 고려함을 통해 맞춤형 공감을 진행하는 것이 공감 대화 도메인에서 긍정적인 결과를 고취시킬 수 있다는 점도 있다.

2. 관련 연구: 공감 대화 생성

EmpatheticDialogues 데이터셋이 공개됨에 따라 [2], 많은 연구들이 소셜 대화 도메인에서 공감을 이해하고 표현할 수 있는 생성 모델들을 제안하였습니다. 예를 들어, mixture of experts 개념을 활용하거나 [3], 대화 상대의 감정을 모방하거나 [4], 상식 지식 (commonsense) 을 사용하거나 [6], causality 를 적용하거나 [8], Rational Speech Acts (RSA) 프레임워크 [5] 를 사용합니다. 또한, 최근 연구 [7]에서는 GPT-3 [16] 가 제로샷/퓨샷 설정에서 정서적 상황 정보를 기반으로 한 in-context example 선택 방법을 사용하여 EmpatheticDialogues 데이터셋에서 Blender 90M [17] 보다 더 나은 성능을 보였다고 보고했습니다. 본 연구는 한국어 공감 대화에서 다양한 대규모 언어 모델들의 성능을 비교 분석하는 첫번째 연구라는 점에서 기존 연구들과 차별점이 있습니다.

3. 방법론

본 장에서는 한국어 공감 대화 데이터셋을 구축한 방법에 대해 설명한다.

3.1 KorEmpatheticDialogues

본 연구에서는 EmpatheticDialogues 데이터셋을 활용하여 한국어 공감 대화 데이터셋을 구축하고자 하며, 이를 위해

Prompt Template for KorEmpatheticDialogues:
아래의 대화에서 비서는 사용자의 상황, 감정, 생각에 잘 공감해야 합니다.

[대화]
`[dialogue]`

Question(질문): 비서의 다음 응답으로 가장 적절한 것은 무엇인가요?

Prompt Template for EmpatheticDialogues
The following is a dialogue with an empathetic assistant. The assistant should empathize well with human's situation, feelings, and thoughts. The dialogue is provided line-by-line.

Dialogue:
`[dialogue]`

그림 1. KorEmpatheticDialogues 데이터셋에서의 언어 모델의 공감 대화 성능 측정 위한 프롬프트 템플릿 (**top**). EmpatheticDialogues 데이터셋에서의 언어 모델의 공감 대화 성능 측정 위한 프롬프트 템플릿 (**bottom**). `[dialogue]` 에는 실제 대화가 입력으로 제공된다.

DeepL 번역기 [1]를 활용하였다. 주의할 점은 대화는 문맥적인 정보를 지니고 있어 이전 문맥 정보를 보지 않고서는 추론하기 어려운 정보들이 존재한다. 그러므로, 문장 단위로 번역을 진행하지 않고 대화 문맥 전체를 번역기에 통과시켜 대화 문맥의 일관성을 유지하고자 하였다. 표 2에서 알 수 있듯이, 구축된 KorEmpatheticDialogues 데이터셋은 총 24,847 의 대화를 포함하며 평균 4개의 발화를 포함하고 있다. EmpatheticDialogues 데이터셋과 마찬가지로, KorEmpatheticDialogues 데이터셋은 대화 마다 하나의 상황과 감정 정보를 포함하고 있으며 KorEmpatheticDialogues 데이터셋의 일부 예시는 표 1와 같다.

3.2 Personality-aware Prompting

본 연구에서는 (1) 다양한 언어 모델들의 공감 대화 능력을 분석하는 것과 (2) MBTI 와 같은 개인 성향이 공감 대화 생성에 끼치는 영향에 대해 분석하는 것을 실험적으로 분석하고자 한다. (1) 을 위해, 그림 1에서 보여지듯이, 대규모 언어 모델의 입력으로 사용될 심플한 프롬프트를 구축하였으며, 한국어/

[1] `https://www.deepl.com/translator`

표 3. MBTI 타입 별 대표 문장

Type	Personality Sentence
F	나는 결정을 내릴 때 상대방의 감정을 중요하게 고려한다.
T	나는 결정을 내릴 때 논리와 분석을 중요하게 고려한다.
E	나는 사람들과 함께 있을 때 에너지를 얻고, 사회적 상황을 즐긴다.
I	나는 혼자 시간을 보내는 것을 선호하고, 사회적 상황에서는 에너지를 소모한다.
S	나는 구체적이고 현재 지향적인 정보를 중시하며, 경험을 통해 세상을 이해한다.
N	나는 미래의 가능성과 추상적인 아이디어에 관심이 많으며, 직관적으로 세상을 이해한다.
P	나는 유연하게 새로운 정보와 가능성에 대응하며, 계획보다는 즉흥적으로 행동한다.
J	나는 계획을 세우고 일을 체계적으로 처리하는 것을 선호한다.

영어 버전의 프롬프트를 개별적으로 구축하였다. (2) 을 위해, 성향을 프롬프트에 반영시키게 유도하는 프롬프트를 그림 x 처럼 한국어 버전만 구축하였다. 초기 실험에서 각 성향에 대한 타입 정보만을 주었을 때 (예: F, T 등) 생성되는 응답의 품질이 떨어지는 것을 확인하였다. 또한, GPT-4 처럼 많은 양의 사전지식을 이해하고 있어서 성향에 대한 정보를 잘 이해하고 있는 것과 달리, 상대적으로 성향에 대해 깊이 있는 이해를 하지 않을 수 있다고 생각하였다. 이에 따라, MBTI 타입 정보를 나타내는 대표 문장을 하나씩 생성하여, 이 문장들을 토대로 언어 모델에게 성향을 주입하고자 하였다. 표 3에서 확인할 수 있다.

4. 실험

4.1 실험 세팅

Datasets. 실험에 사용된 데이터셋은 한국어/영어 공감 대화 데이터셋으로 총 2가지이며, 대규모 언어 모델들의 공감 대화 능력 성능을 측정하는 것이 주된 목표이므로 따로 학습 데이터셋을 사용하지 않고 테스트셋에 대해서만 실험을 진행한다. 또한, 본 연구에서는 OpenAI API 비용 절감 및 여러 언어 모델들 간의 공감 대화 능력의 빠른 비교를 위해 전체 테스트셋 중 랜덤으로 100 건을 샘플링하여 비교 실험을 위해 사용한다.

Large Language Models. 본 실험에 사용된 대규모 언어 모델의 종류는 다음과 같다. 먼저, 한국어 언어 모델로는 KoAlpaca 5.8B 와 KoAlpaca 12.8B 를 사용하였으며, 다국어 언어 이해 능력도 있는 ChatGPT 와 GPT-4 도 사용하였다. 다음으로, 영어 공감 대화 생성을 위해서는 다양한 데이터셋 및 다양한 방법으로 학습된 언어 모델들을 사용하였다. 한국어 공감 대화 실험과 마찬가지로 ChatGPT 와 GPT-4 를 영어 공감 대화 생성 능력 평가를 위해 사용한다. 그리고, 오픈 소스 언어 모델로는, Alpaca 13B, Vicuna 13B, WizardLM 13B, Tulu 13B, Code-Alpaca, FLAN V2, ShareGPT, Self-Instruct, OpenAssistant, Dolly V2, GPT-4-Alpaca 를 실험에 사용하였다. 그리고, 언어 모델의 사이즈에 따른 성능 비교를 위해 LLaMa2 Chat 7B, 13B, 70B 에 대해서도 실험을 진행하였다.

Implementation Details. 모든 실험은 총 8대의 A100 GPU (40GB) 에서 이루어졌으며, 모든 언어 모델의 생성과정에 있어서, temperature 는 0.9 로 설정하였고, 최대 생성 길이를 1024 로 제한하였다.

Evaluation Metrics. 공감 대화 생성에서 가장 큰 문제점은 정량적 평가를 하기가 어렵다는 것이다. 그러한 이유로는, 공감이라는 것은 사용자들의 성향에 따라 다르게 느낄 여지가 있는 주관성을 띄기 때문이다. 따라서, 최대한 holistic 한 평가를 진행하기 위해, 기존 연구 [18, 19]에서처럼 GPT-4 를 이용하여 평가를 진행한다. 평가 항목으로는 크게 4가지를 설정하였으며, 1) Empathy: 모델이 생성한 응답이 내담자의 상황을 잘 이해하며 적절히 공감하는가? 2) Explorations (EX): 모델이 생성한 응답이 내담자의 정서적 상황을 잘 파악하려고 시도하는거 같은가? 3) Interpretations (IP): 모델이 생성한 응답이 내담자의 정서적 상황을 잘 이해하고 있는가? 4) Emotional Reactions (ER): 모델이 생성한 응답이 내담자를 향해 정서적인 반응을 잘 보이고 있는가? 각 항목은 5-Likert scale 에 기반을 하고 있다. 실제 GPT-4 를 활용한 평가를 위해 사용한 프롬프트 템플릿은 그림 2 와 같다.

4.2 실험 결과

Zero-shot Performance. 표 4 는 다양한 언어 모델들의 공감 대화 생성에서 Empathy, Interpretations, Explorations, Emotional Reactions 측면에서 측정한 제로샷 성능을 보여주고 있다. 흥미로운 것은, 기존 연구 [19]와 다르게 오픈 소스 언어 모델인 LLaMa2 Chat 가 모든 지표에서 GPT-4 보다 높은 성능을 보여주고 있다. 또한, human-authored 데이터셋으로 학습된 언어 모델들 (i.e., FLAN V2, Dolly V2, OpenAssistant) 보다 ChatGPT 와 같은 proprietary 언어 모델로부터 확보된 데이터셋 (dataset distillation) 으로 학습된 언어 모델들 (i.e., ShareGPT, GPT-4-Alpaca)의 성능이 보다 높은 것을 확인할 수 있다. 이것이 시사하는 바는 dataset distillation 방법을 통한 언어 모델을 구축하는 최근 방법이 social knowledge reasoning 이 필요한 공감 대화 생성 분야에서도 효과적인 방법이라는 것을 알 수가 있음을 통해, 최근 방법의 효과를 입증한다. 한국어 공감 대화 생성 분야에서는 기존 연구 [19]에서의 결과와 유사하게 아직 ChatGPT 와 GPT-4 모델들과 오픈 소스 언어 모델 간의 성능 차이가 큰 것을 알 수 있다.

Code-tuning is helpful? 표 4 에서 보여지듯이, Code-Alpaca 모델이 code 데이터셋에 튜닝되었지만 공감 대화 도메

You will be given one response for one dialogue.

Your task is to rate the response based on the criteria provided.

Please make sure you read and understand these criteria carefully.

Evaluation Criteria:

Empathy (1-5) - Does the response show understanding of the situation and empathize appropriately?
Explorations (1-5) - Does the response make an attempt to explore the interlocutor's experiences and feelings?
Interpretations (1-5) - Does the response communicate an understanding of the interlocutor's experiences and feelings?
Emotional Reactions (1-5) - Does the response express or allude to warmth, compassion, concern, or similar feelings towards the interlocutor?

[Dialogue]
`[dialogue]`

[Assistant's Response]
`[response]`

[The End of Assistant's Response]

Please give feedback on the assistant's responses. Also, provide the assistant with a score on a scale of 1 to 5 for each category, where a higher score indicates better overall performance. Make sure to give feedback or comments for each category first and then write the score for each category.

Lastly, return a Python dictionary object that has criteria names as keys and the corresponding scores as values.

[System]

그림 2. 공감 대화 생성 평가를 위해 사용한 프롬프트 템플릿

인에서 준수한 성능을 보이고 있다. 심지어는 Self-Instruct, FLAN V2, OpenAssistant 모델보다 높은 성능을 보이고 있다. 이를 통해, code 데이터셋으로 튜닝하는 것이 효과적이라고 생각할 수 있으나, 직접적으로 code 튜닝이 공감 대화 생성에 있어서 효과적이라기 보다는 Code-Alpaca 의 경우에는 이미 Alpaca 13B 모델에서 20K 사이즈의 instruction-following 데이터셋으로 튜닝 시킨 모델이여서, 준수한 성능을 달성한 것은 Alpaca 13B 모델의 성능이 좋았기 때문이라고 볼 수 있다.

Effect of Model Scaling. 모델 사이즈에 따른 공감 대화 성능 변화를 관찰하기 위해, 영어 언어 모델은 LLaMa2 Chat 7B, 13B, 70B 를 선택하였고, 한국어 언어 모델은 KoAlpaca 5.8B, KoAlpaca 12.8B 를 선택하였다. 표 4에서 알 수 있듯이, 한국어 언어 모델의 경우에는 모델 사이즈가 커짐에 따라 제로샷 성능이 일괄적으로 상승하였다. 그러나, 영어 언어 모델의 경우에는 모델 사이즈가 비약적으로 증가하여도 성능이 줄어드는 현상을 관측하였다. 이러한 현상이 발생한 이유로는 평가에 사용된 테스트 셋 샘플의 수가 적은데 비해 해당 데이터셋이 LLaMa2 Chat 에 우호적인 샘플이었을 수 있고, GPT-4 기반의 평가 방법의 verbosity bias 때문일 수 있다. verbosity bias 의 경우에는 GPT-4 이 긴 응답을 상대적으로 짧은 문장보다 더 높은 점수를 부과하는 으로써 [20, 21], 그림 3에서도 나타나듯이 공감 대화 생성 도메인에서도 비슷한 현상이 발생했다. 자세한 건 아래에서 설명한다.

Verbosity Bias. 위에서 언급하였듯이, 대규모 언어 모델을 평가 지표로 사용하는 추세가 최근에 많은 연구들 [19]에서 보여지고 있으나 여러 문제점들이 보고 되고 있다. 그 중에, 대규모 언어 모델들이 긴 응답에 더 높은 점수를 매기는 현상인 verbosity bias 가 있다 [20, 21]. 그림 3에서 보여지듯이, 전반적으로 언어 모델이 생성하는 응답의 길이가 길어짐에 따라 더 높은 성능을 기록하는 것을 확인할 수 있다. 특히, 표 4에서 가장 높은 성능을 기록한 LLaMa2 Chat 모델들이 상당히 긴 응답을 생성하고 있으며 이에 따라 가장 높은 성능을 기록하였다. 이러한 현상은 verbosity bias 로 볼 수 있으며, 보다 적합한 평가를 위해 실제 사용자 평가를 하는 것이 필요하다는 것을

표 4. 여러 대규모 언어 모델들의 공감 대화 생성 분야에서의 제로샷 성능 비교

Models	Empathy ↑	EX ↑	IP ↑	ER ↑
Zero-shot performance on EMPATHETICDIALOGUES				
self-instruct-13b	$2.05_{\pm 1.52}$	$1.33_{\pm 1.19}$	$1.89_{\pm 1.59}$	$1.8_{\pm 1.48}$
flan-v2-13b	$2.74_{\pm 1.0}$	$1.86_{\pm 1.18}$	$2.52_{\pm 1.28}$	$2.21_{\pm 1.11}$
oasst1-13b	$3.15_{\pm 1.79}$	$2.44_{\pm 1.56}$	$3.17_{\pm 1.86}$	$2.99_{\pm 1.79}$
alpaca-13b	$3.79_{\pm 1.52}$	$2.35_{\pm 1.18}$	$3.75_{\pm 1.53}$	$3.65_{\pm 1.55}$
code-alpaca-13b	$3.16_{\pm 1.56}$	$1.95_{\pm 1.09}$	$3.06_{\pm 1.61}$	$2.9_{\pm 1.58}$
dolly-13b	$3.57_{\pm 1.14}$	$2.0_{\pm 0.75}$	$3.46_{\pm 1.24}$	$3.28_{\pm 1.32}$
gpt4-alpaca-13b	$4.26_{\pm 1.22}$	$3.38_{\pm 1.47}$	$4.32_{\pm 1.19}$	$4.07_{\pm 1.36}$
sharegpt-13b	$4.31_{\pm 1.04}$	$3.2_{\pm 1.38}$	$4.26_{\pm 1.11}$	$4.14_{\pm 1.25}$
tulu-13b	$4.06_{\pm 1.02}$	$3.13_{\pm 1.4}$	$4.07_{\pm 1.08}$	$3.93_{\pm 1.16}$
vicuna-13b	$4.63_{\pm 0.8}$	$3.32_{\pm 1.15}$	$4.57_{\pm 0.91}$	$4.49_{\pm 1.0}$
wizard-13b	$4.68_{\pm 0.65}$	$3.48_{\pm 1.09}$	$4.7_{\pm 0.57}$	$4.53_{\pm 0.82}$
llama-2-7b-chat	$4.96_{\pm 0.2}$	$4.66_{\pm 0.76}$	$4.98_{\pm 0.14}$	$4.95_{\pm 0.22}$
llama-2-13b-chat	$4.97_{\pm 0.17}$	$4.62_{\pm 0.86}$	$4.97_{\pm 0.17}$	$4.96_{\pm 0.2}$
llama-2-70b-chat	$4.94_{\pm 0.28}$	$4.52_{\pm 0.81}$	$4.96_{\pm 0.2}$	$4.92_{\pm 0.31}$
gpt-3.5-turbo	$4.66_{\pm 0.53}$	$3.34_{\pm 1.17}$	$4.72_{\pm 0.49}$	$4.5_{\pm 0.69}$
gpt-4	$4.75_{\pm 0.48}$	$3.19_{\pm 1.06}$	$4.8_{\pm 0.42}$	$4.65_{\pm 0.67}$
Zero-shot performance on KOREMPATHETICDIALOGUES				
koalpaca-polyglot-5.8B	$1.35_{\pm 0.64}$	$1.34_{\pm 0.79}$	$1.32_{\pm 0.66}$	$1.26_{\pm 0.58}$
koalpaca-polyglot-12.8B	$1.97_{\pm 1.08}$	$2.01_{\pm 1.2}$	$1.94_{\pm 1.15}$	$1.81_{\pm 0.99}$
gpt-3.5-turbo	$4.5_{\pm 0.5}$	$4.5_{\pm 0.5}$	$4.5_{\pm 0.5}$	$4.0_{\pm 1.0}$
gpt-4	$4.0_{\pm 1.0}$	$2.0_{\pm 1.0}$	$4.0_{\pm 1.0}$	$4.5_{\pm 0.5}$

나타낸다.

Few-shot Performance. 그림 4 은 영어 공감 대화 데이터셋에서 CHATGPT 와 GPT-4 의 퓨샷 성능을 보여주고 있다. 전반적으로, EX 지표를 제외하곤 퓨샷을 적용하였을 때 일괄적으로 성능이 감소하였으며, 이러한 현상은 기존의 연구들에서 보여지는 양상과 다르다. 이러한 이유로는 대규모 언어 모델을 평가 지표로써 사용할 때 발생하는 문제점이 원인일 수도 있고, 퓨샷 샘플의 여러 sensitivity 문제일 수도 있다. 예를 들어, 퓨샷 샘플의 순서에 따른 영향 [22] 혹은 정서적 상황 불일치 [7] 가 있을 수 있다. 따라서, 보다 정밀하고 면밀한 퓨샷 성능을 관측하기 위해서는 정교한 설계가 필요하다고 생각된다.

Effect of Personality. 표 4는 MBTI 타입 별 공감 대화의 제로샷 성능이 어떻게 달라지는지를 보여주고 있다. 실험을

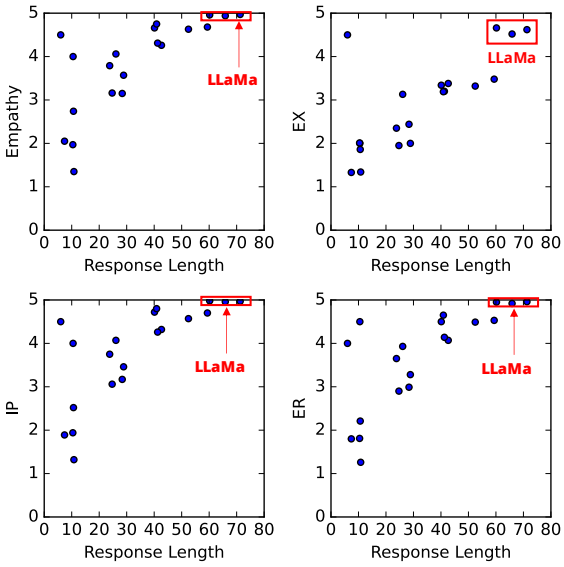

그림 3. 언어 모델이 생성하는 응답의 길이와 공감 대화 성능의 상관관계 비교 분석. 각 파란색 점은 표 4에 있는 모델들을 나타냄.

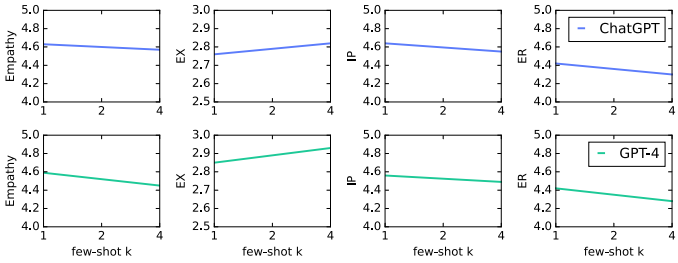

그림 4. CHATGPT 와 GPT-4 모델의 EMPATHETICDIALOGUES 에서의 공감 대화 퓨샷 성능 비교

위해 사용한 모델은 KOALPACA 12.8B 이며, 타입 별로 20개의 대화씩 총 160개의 대화 샘플에 대해 실험을 진행했다. 각 차원에서 비교를 하면, F, P, I, N 타입이 공감 대화 생성에서 상대적으로 높은 성능을 기록하였다. 흥미로운 사실은, T 보다는 F 의 성향이 주입된 프롬프팅 방법이 공감 대화 성능이 높았으며, 이는 현재 사회 전반에 만연하는 "F들은 공감을 더 잘한다" 와 일관된다. 또한, 표 3에서 구축한 personality 대표 문장이 언어 모델에게 personality-aware 응답을 생성한 것을 확인할 수 있다. 향후에는 간접적으로 프롬프팅을 통해 성향을 주입하기 보다는 기존 연구 [23]에서처럼 직접적으로 모델 튜닝을 통해 주입하고자 한다.

Case Study. 표 6는 MBTI 유형에 따른 KOALPACA 12.8B 모델의 다른 생성 결과를 보여준다. 주어진 대화 문맥에 대해 personality 정보가 없는 zero-shot in-context learning (ICL) 의 경우에는 도움을 주려는 경향의 응답을 생성하는 것을 알

표 5. MBTI 타입에 따른 공감 대화 생성에서 제로샷 성능 비교

MBTI Type	Empathy ↑	EX ↑	IP ↑	ER ↑
F	$1.92_{\pm 0.98}$	$2.06_{\pm 1.14}$	$1.85_{\pm 0.97}$	$1.75_{\pm 0.88}$
T	$1.9_{\pm 1.16}$	$1.96_{\pm 1.21}$	$1.84_{\pm 1.16}$	$1.68_{\pm 0.97}$
J	$1.84_{\pm 1.12}$	$2.03_{\pm 1.21}$	$1.76_{\pm 1.08}$	$1.66_{\pm 0.95}$
P	$1.99_{\pm 1.02}$	$2.0_{\pm 1.14}$	$1.92_{\pm 1.05}$	$1.76_{\pm 0.9}$
E	$1.81_{\pm 1.04}$	$1.97_{\pm 1.15}$	$1.72_{\pm 0.99}$	$1.65_{\pm 0.86}$
I	$2.01_{\pm 1.09}$	$2.05_{\pm 1.2}$	$1.96_{\pm 1.12}$	$1.78_{\pm 0.99}$
N	$1.94_{\pm 1.08}$	$1.99_{\pm 1.23}$	$1.89_{\pm 1.1}$	$1.7_{\pm 0.89}$
S	$1.89_{\pm 1.06}$	$2.04_{\pm 1.12}$	$1.8_{\pm 1.03}$	$1.72_{\pm 0.96}$

표 6. KoAlpaca 12.8B 모델의 생성 결과

Emotion: Lonely

Situation: 8년 동안 사귀었던 여자친구와 방금 헤어졌는데 너무 허전해요.

Dialogue:

발화자 1: 나는 거기에, 무엇을해야할지 모르겠다, 내 여자 친구와 헤어졌고, 우리는 8 년 동안 함께했다.

발화자 2: 유감입니다! 이별에 대해 조금이라도 알고 있나요? 그것에 대해 생각 했습니까?

발화자 1: 네, 우리는 마음으로 함께 결정했고 집에 돌아와서 세상과 너무 멀어진 것을 알고 있습니다.

Golden Response: 다시 한번 죄송합니다! 이 슬픔에서 벗어날 수 있기를 바랍니다. 이 슬픔에서 벗어나기 위해 관심사에 집중해 주세요.

ICL Response:

거기서 어떤 종류의 도움이 필요하신가요?

ENFJ Response:

어떤 위로의 말씀을 드려야 할지 잘 모르겠습니다. 이전에 이별을 경험 한 적이 있나요?

ENTJ Response:

이별에 대해 어떻게 느끼셨나요? 어떤 유형의 도움이 필요한가요?

수 있다. 이는 최근의 RLHF 로 학습된 언어 모델들에 사용된 human alignment 데이터셋의 특성으로 인해 기인한 것이라고 볼 수 있다. MBTI 성향이 주입된 프롬팅의 경우에는 MBTI 유형에 상응되는 결과를 보여주고 있다. ENFJ 가 상대적으로 내담자의 상황을 이해하여 정서적인 공감을 시도하려는 것을 볼 수 있으며, 그에 반해 ENTJ 는 논리와 분석에 좀 더 집중한 형태의 응답을 생성하는 것을 확인할 수 있다.

5. 논의 및 한계점

본 연구에서는 다양한 언어 모델들의 한국어/영어 공감 대화에서의 성능 비교와 personality와 공감 사이에서의 상관관계를 실험적으로 분석하였다. 그러나, 몇 가지 문제점이 있다.

Quality Inspection. 번역기를 통해 자동으로 구축된 KorEmpatheticDialogues 데이터셋의 품질을 평가하는 것이 필요로 하다. 왜냐하면, 영어와 한국어 사이에서는 사회적, 문화적 차이가 존재하고 이런 부분들이 번역기에는 반영되어 있지 못하다. 또한, 번역기는 실제 일상생활 대화체를 반영하여 번역하고 있지 않으므로, 번역된 문장은 대부분 문어체여서 일상생활 대화에 자주 등장하는 구어체와는 차이가 있을 수 있다. 그러므로, 다른 연구자들도 이용할 수 있게끔 하기 위해서는 기존 연구 [24]처럼 사용자가 직접 저품질의 샘플을 수정하거나 사용자 평가를 거치는 과정이 필요하다.

Lack of Human Evaluator. 본 연구의 실험에서 평가 지표로 GPT-4 를 사용하였다. 그러나, 언어 모델 기반의 평가 지표는 여전히 여러 문제들이 있으며, 이를 보완하기 위해 사용자 평가 실험을 진행할 필요가 있다. 사용자 평가를 진행할 때에는 개별 평가자의 성향을 수집한 후에 성향과 공감 사이의 상관관계를 분석하게 된다면 보다 깊은 분석 결과를 얻을 수 있을 것이다.

6. 결론

본 연구에서는 대규모 언어 모델들의 한국어/영어 공감 대화 생성에서의 성능 비교를 다양한 지표에서 진행하였으며, 이를 위해 번역기를 통해 KorEmpatheticDialogues 데이터셋을 구축하였다. 또한, personality 와 공감 사이의 상관관계를 분석하기 위해 personality-aware prompting 방법론을 제안하였다. 포괄적인 실험을 통해, 공감 대화 생성 분야에서 어떤 부분들을 고려하여 설계하는 것이 좋을지에 대해 실험적으로 보여준다. 향후 연구로는 RLHF 를 확장하여 personality-alignment 를 학습한 대규모 언어 모델을 구축하고자 한다.

감사의 글

이 논문은 2023년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: 확장현실(XR) 기반 복합테러 대응 교육·훈련 테스트 베드 구축 / 연구개발과제번호: PR08-04-000-21]

참고문헌

[1] M. Sap, R. LeBras, D. Fried, and Y. Choi, "Neural theory-of-mind? on the limits of social intelligence in large lms," *arXiv preprint arXiv:2210.13312*, 2022.

[2] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[3] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," *arXiv preprint arXiv:1908.07687*, 2019.

[4] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, "Mime: Mimicking emotions for empathetic response generation," *arXiv preprint arXiv:2010.01454*, 2020.

[5] H. Kim, B. Kim, and G. Kim, "Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes," *arXiv preprint arXiv:2109.08828*, 2021.

[6] S. Sabour, C. Zheng, and M. Huang, "Cem: Commonsense-aware empathetic response generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 10, pp. 11 229–11 237, 2022.

[7] Y.-J. Lee, C.-G. Lim, and H.-J. Choi, "Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation," *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 669–683, 2022.

[8] J. Wang, Y. Cheng, and W. Li, "Care: Causality reasoning for empathetic responses by conditional graph generation," *arXiv preprint arXiv:2211.00255*, 2022.

[9] I. B. Myers, "The myers-briggs type indicator: Manual (1962)." 1962.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730–27 744, 2022.

[11] OpenAI, "ChatGPT," https://openai.com/blog/chatgpt/, 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[14] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," *arXiv preprint arXiv:2212.08061*, 2022.

[15] S. R. Moghaddam and C. J. Honey, "Boosting theory-of-mind performance in large language models via prompting," *arXiv preprint arXiv:2304.11490*, 2023.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[17] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith *et al.*, "Recipes for building an open-domain chatbot," *arXiv preprint arXiv:2004.13637*, 2020.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "Gpteval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[19] S. Ye, D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, and M. Seo, "Flask: Fine-grained language model evaluation based on alignment skill sets," *arXiv preprint arXiv:2307.10928*, 2023.

[20] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *arXiv preprint arXiv:2306.05685*, 2023.

[21] P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, "Large language models are not fair evaluators," *arXiv preprint arXiv:2305.17926*, 2023.

[22] E. Perez, D. Kiela, and K. Cho, "True few-shot learning with language models," *Advances in neural information processing systems*, Vol. 34, pp. 11 054–11 070, 2021.

[23] E. Choi, Y. Jo, J. Jang, J. Jang, and M. Seo, "Fixed input parameterization for efficient prompting," *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8428–8441, 2023.

[24] J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh, "Kornli and korsts: New benchmark datasets for korean natural language understanding," *arXiv preprint arXiv:2004.03289*, 2020.