Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation

Dongjin Kang[1]* **Sunghwan Kim**[1]* **Taeyoon Kwon**[1] **Seungjun Moon**[1]
Hyunsouk Cho[2] **Youngjae Yu**[1] **Dongha Lee**[1] **Jinyoung Yeo**[1]
[1]Yonsei University [2]Ajou University
{hard1010,kimsh8564,yjy,donalee,jinyeo}@yonsei.ac.kr
hyunsouk@ajou.ac.kr

ACL 2024

Abstract

Strategies	Stages			Examples
Question				*Can you talk more about your feelings at that time?*
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*
Reflection of Feelings				*I understand how anxious you are.*
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*
Affirmation and Reassurance				*You've done your best and I believe you will get it!*
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*
Others				*I am glad to help you!*

- Emotional Support Conversation (ESC)

  : alleviating individuals' emotional distress through daily conversation

- 최근 LLM의 뛰어난 대화 능력에도 불구하고, useful emotional support를 제공하는 데는 한계가 있음

- 이 논문에서는 ESConv데이터셋을 사용하여 LLM의 생성 결과들을 분석함

  : correct strategy를 선택하는데 한계가 있고, 특정 strategy를 선호하는 경향성

  : LLM의 inherent preference가 emotional support에 미치는 영향을 분석

  : 결과적으로, 특정 strategy를 선호하는 경향이 효과적인 emotional support를 방해하고, 적절한 strategy를 선택하는데 부정적인 영향을 준다는 것을 보임

- LLM이 proficient emotional supporter 역할을 위해 필요한 approach들에 대한 insight를 제공

Introduction

- Effective emotional support

  : helpful emotional support를 제공하는 것 뿐만 아니라,

  psychological, relational, and physical problem을 악화시킬 수 있는 poor-quality emotional support를 피해야 함

  → 하지만 이러한 emotional support를 제공한다는 것은 매우 복잡하고 심지어 human도 여전히 challenging

- 최근 LLMs의 remarkable conversational ability로, 다양한 dialogue system에서 사용되고 있음

  : 특히, LLM을 사용해서 professional counseling이 아닌 daily conversation에서 emotional support를 제공하고자 하는 연구들이 있었음

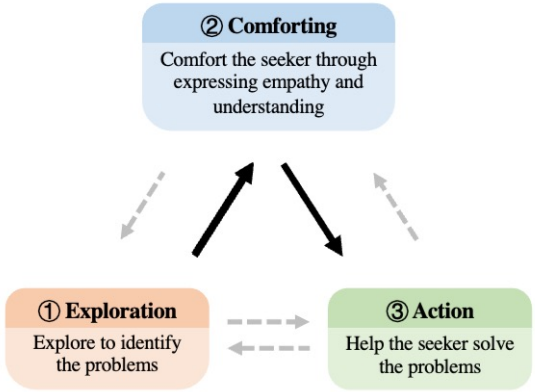
  → 하지만 outstanding capabilities의 LLM도 emotional support에는 한계가 있었음

Introduction

- Hill's helping skill theory를 기반으로, emotional support를 위한 framework와 데이터셋을 제안하는 연구도 있었음.
  : 3가지의 stage, 총 8개의 support strategies로 구성되어 있음

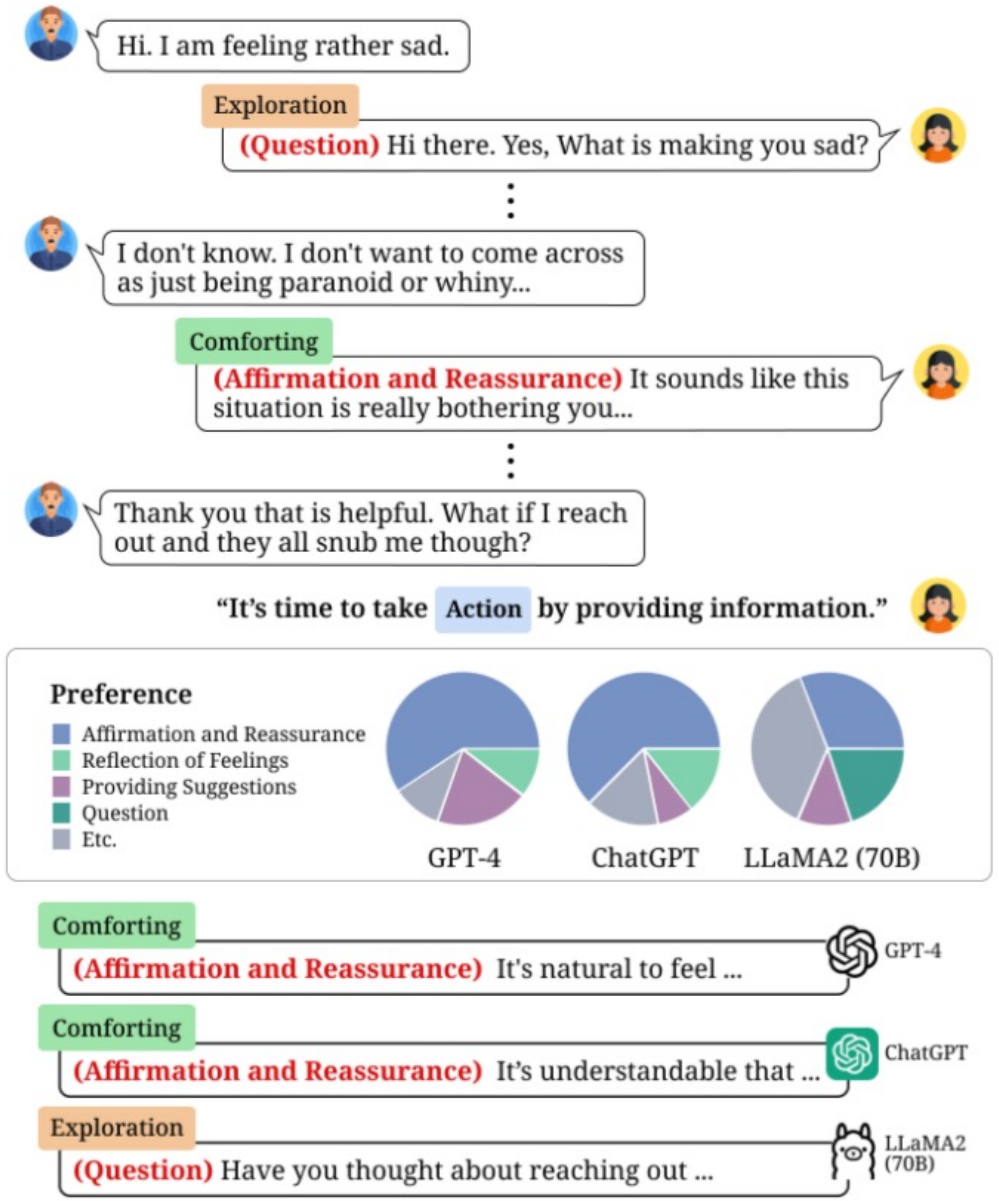
- ESC task는 보통 두 단계로 구성됨 : strategy selection and strategy-constrained response generation
  : appropriate strategy를 선택하는 것이 중요함

- LLM이 accurate strategy를 예측하는데 한계가 있음을 발견하고, 그 이유를 발견하고자 함
  : 얼마나 자주 LLM이 각 strategy를 선택하는지 분포를 측정하고, 특정 strategies에 대한 high preference를 확인함



Strategies	Stages			Examples
Question				*Can you talk more about your feelings at that time?*
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*
Reflection of Feelings				*I understand how anxious you are.*
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*
Affirmation and Reassurance				*You've done your best and I believe you will get it!*
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*
Others				*I am glad to help you!*

Introduction



Introduction

- RQ1: Does the preference affect providing emotional support?

  다양한 LLM의 proficiency를 측정하여 각 모델이 잘 예측하거나 잘 예측 못하는 strategies와 stages를 확인함

  → 모델이 preference가 높은 strategy를 사용하거나 preference가 높은 strategy가 속한 stage의 strategy를 사용했을 때 향상된 성능을 보임을 증명함

  → robust한 prediction을 위해서는 이러한 preference 편향성을 낮추는 것이 중요하다

- RQ2: How to mitigate the preference bias on LLMs?

  LLM이 Contact Hypothesis 이론과 일치한다는 것을 발견

  → LLM의 preference bias를 줄이기 위해서는 external assistance가 필요하다고 주장

  → preference bias를 줄였을 때, 3 stages 모두에서 일관된 strategy 예측 성능을 보임

* 접촉 가설(Contact Hypothesis)
: 집단 간 편견을 감소시키기 위해 가장 포괄적으로 연구되어 온 대표적인 심리학 이론.
: 집단간의 접촉을 통해 집단간의 편견과 차별을 해소하고 관계를 개선할 수 있다는 이론

Introduction

- RQ3: Does improving preference bias indeed help to become a better emotional supporter?

  response가 helpful emotional support를 지원하는지 평가하기 위해 공식적인 criteria를 구축

  → criteria를 기반으로 하는 human evaluation에서 낮은 preference bias일 수록 높은 점수를, 높은 preference bias일수록 낮은 점수를 받음

Evaluation Setup – Task and Focus

- Task: emotional support response generation

  dialogue background $\mathcal{I}$(situation), dialogue context C → strategy S

  dialogue background $\mathcal{I}$(situation), dialogue context C, strategy S → response R

$$\mathcal{S} \sim P_\theta(\cdot|\mathcal{I}, \mathcal{C}) \quad (1)$$

$$\mathcal{R} \sim P_\theta(\cdot|\mathcal{I}, \mathcal{C}, \mathcal{S}) \quad (2)$$

- Focus: strategy-centric analysis

  LLM이 emotional support를 수행하지 못하는 다양한 이유들이 있지만, 이 논문에서는 **strategy**에만 focus

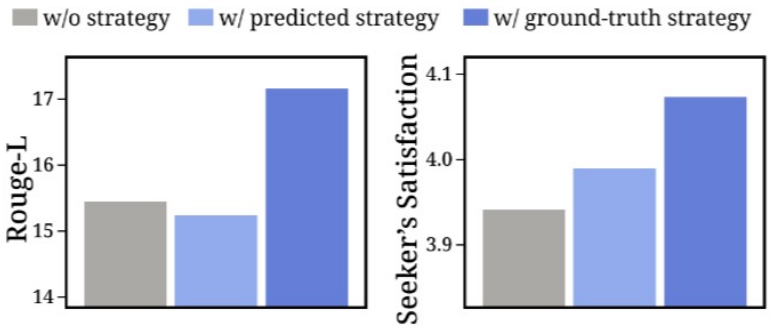


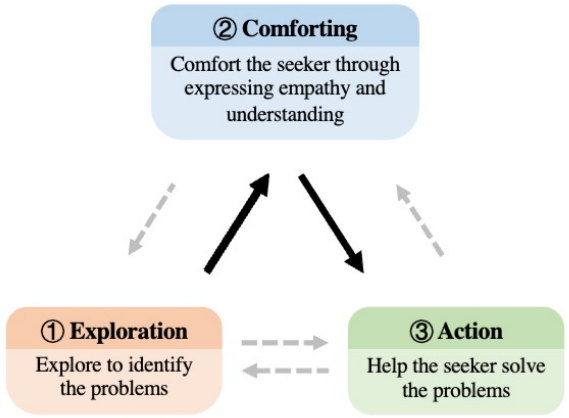
Figure 2: The results of strategy-constrained responses on both automated and human evaluation, showing the efficacy of strategy on ChatGPT. Appropriate strategy significantly enhances the quality of emotional support responses. The details are in Appendix A.2.

Evaluation Setup – Evaluation Set

- three test sets D_t

Strategy	Exploration D_1	Comforting D_2	Action D_3	**Total** (D)
Que.	**24.8**	10.0	7.0	12.8
Res.	**16.8**	9.6	4.5	9.4
Ref.	**16.8**	**18.3**	6.3	12.7
Sel.	**16.8**	**20.1**	**15.4**	17.2
Aff.	7.6	**24.1**	**21.1**	18.2
Pro.	8.4	8.5	**24.4**	15.3
Inf.	6.5	6.5	**18.5**	11.7
Oth.	2.3	2.5	2.8	2.6

Table 1: The ratio (%) of support strategies in our test sets. Each test set D_t is composed with samples corresponding to each stage. The highlighted strategies are primarily utilized in each stage (Liu et al., 2021).



Strategies	Stages	Examples
Question		*Can you talk more about your feelings at that time?*
Restatement or Paraphrasing		*It sounds that you feel like everyone is ignoring you. Is it correct?*
Reflection of Feelings		*I understand how anxious you are.*
Self-disclosure		*I feel the same way! I also don't know what to say to strangers.*
Affirmation and Reassurance		*You've done your best and I believe you will get it!*
Providing Suggestions		*Deep breaths can help people calm down. Could you try to take a few deep breaths?*
Information		*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*
Others		*I am glad to help you!*

Evaluation Setup – Metrics

- We propose a new suite of metrics that focus on strategies: proficiency, preference, and preference bias.

1) Proficiency

how well the model selects the correct strategy

전체 test 데이터셋에 대한 macro F1 score

각 test 데이터셋에 대한 weighted F1 score

2) Preference

how much the model prefers certain strategies over others

각 strategy에 대한 선호도를 측정하기 위해 Bradley-Terry model 적용

Strategy i에 대한 preference $p_i' = \frac{\sum_j (w_{ij} p_j)/(p_i + p_j)}{\sum_j w_{ji}/(p_i + p_j)}$ w_{ij} : gt가 j일때 모델이 i를 예측한 횟수

모든 p_i는 1로 initialize되어 있고 iteration을 통해 update되어 나감,
즉 $p`_i$는 다음 iteration에서의 i에 대한 preference

Evaluation Setup – Metrics

- We propose a new suite of metrics that focus on strategies: proficiency, preference, and preference bias.

3) Preference Bias

preferences p_i 의 표준편차를 preference bias로

→ 값이 클수록 preferred and non-preferred strategies에 대한 선호도가 명확하다

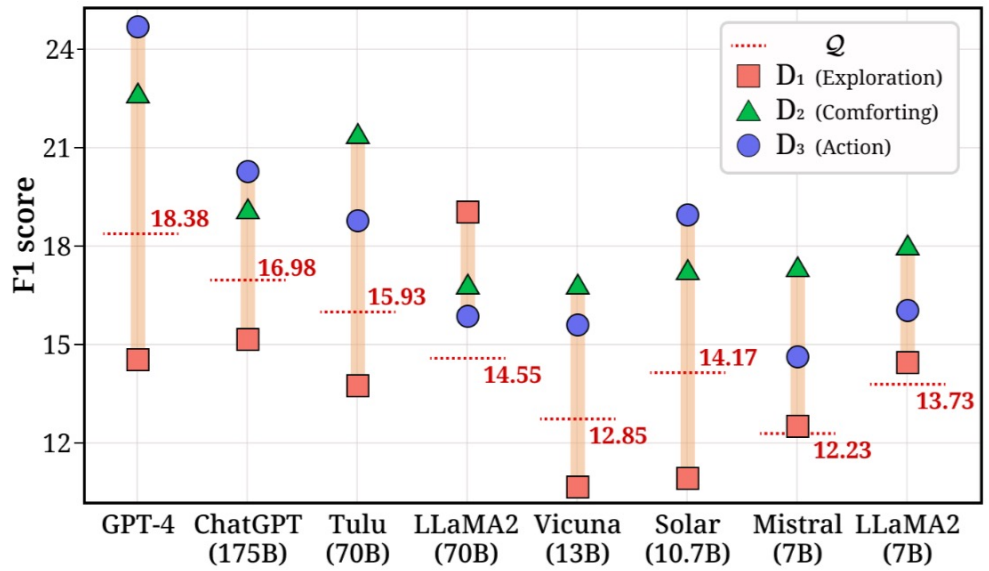
Models & Implementation Details

- Closed-source models
  : ChatGPT and GPT4
- Open-source models
  : LLaMA2-7B/70B, Tulu70B, Vicuna-13B, Solar-10.7B and Mistral7B

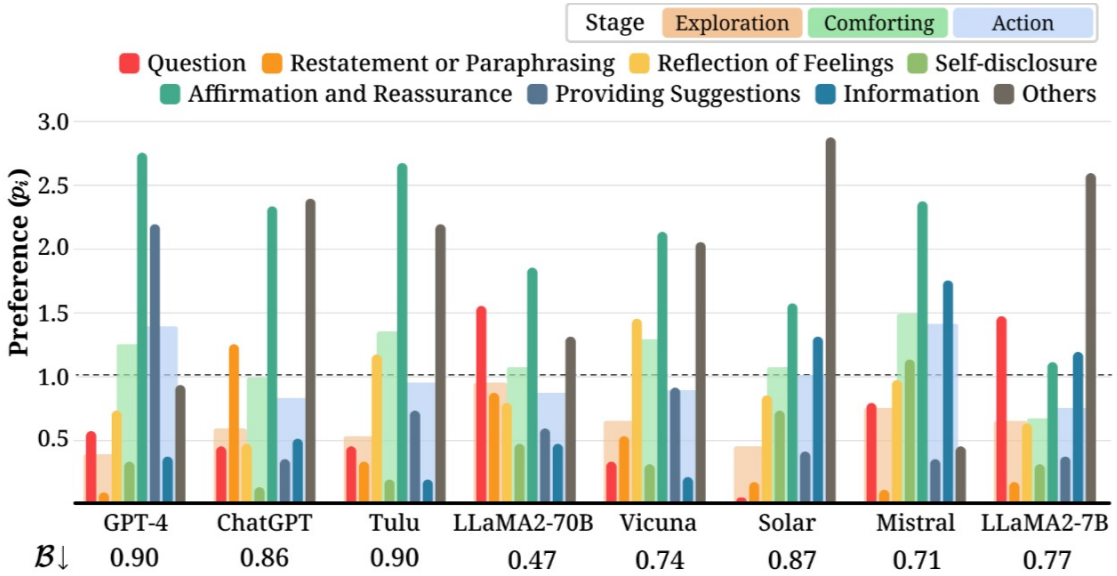
- Prompt
  : strategy descriptions
  : 2-shot examples

RQ1: Does the preference affect providing emotional support?

- Proficiency of LLMs



(a)



(b)

- 큰 모델일수록 좋은 성능, 작은 모델일수록 낮은 성능을 보이는 경향성
- 각 testset별 성능에서는 대부분 D2와 D3가 더 좋은 성능을 보임

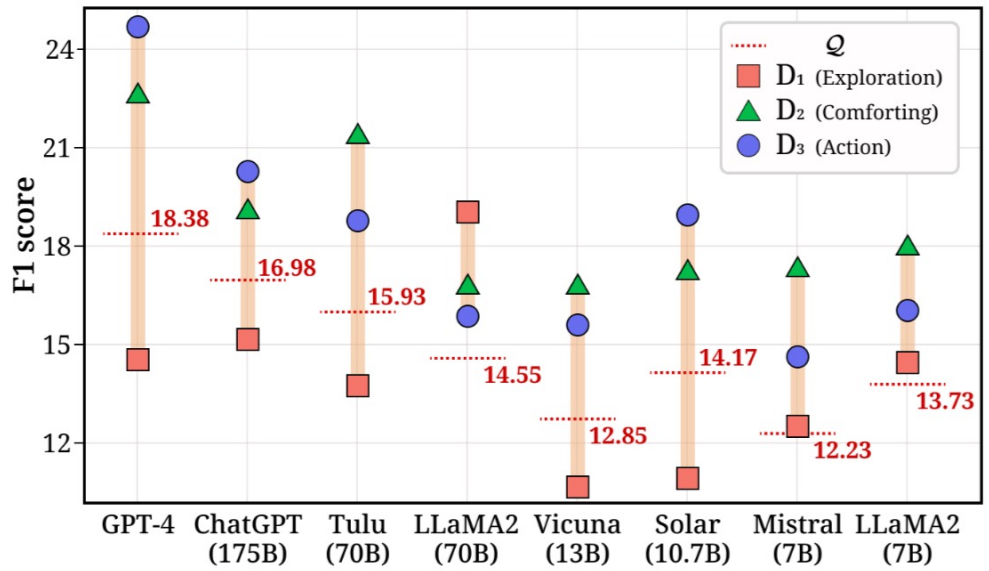
: LLM들이 일반적으로 comforting or action과 같은 답변은 잘 생성하지만, exploration stage에서는 poor-quality emotional support를 제공

→ Stage 구성이 Exploration → Comforting→ Action 이기 때문에 첫번째 stage에서 poor-quality response를 제공하게 되므로, 최종적으로 effective emotional support를 제공하기 어려울 수 있음

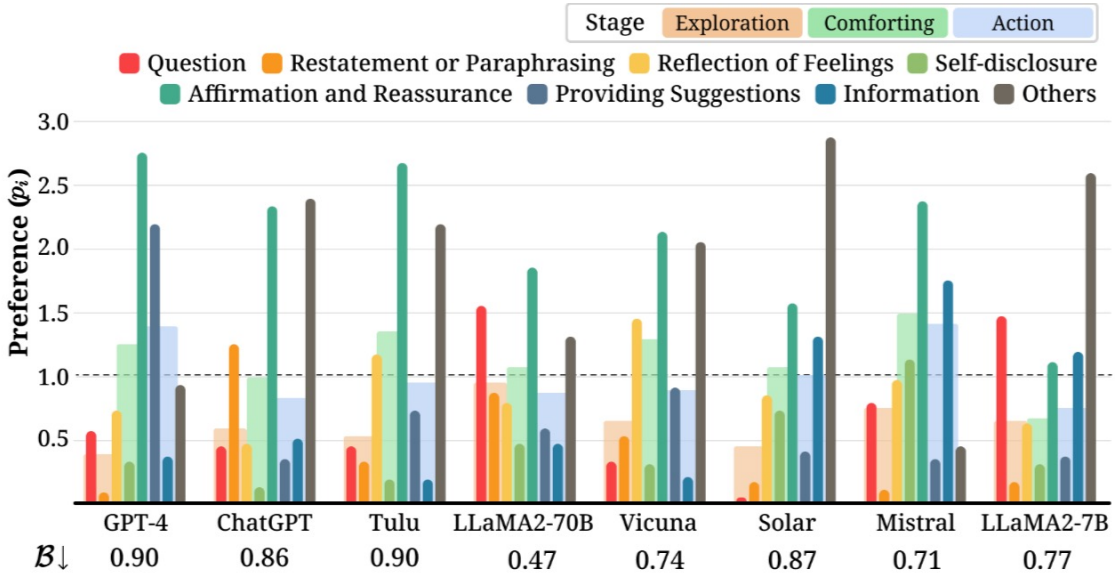
→ Macro F1 성능이 높다고해서 반드시 helpful emotional support를 제공한다고 할 수 없음

RQ1: Does the preference affect providing emotional support?

- Proficiency of LLMs



(a)



(b)

- 각 모델별로 Stage에 대한 preference가 높으면 해당 stage의 성능도 높은 것을 알 수 있음
- llama2의 경우 각 stage별로 균일한 preference를 보이고, D_t에서도 robust한 성능을 보임

→ Macro F1에서 좋은 성능을 보여도, significant preference bias가 특정 stage에서 낮은 성능을 야기할 수 있음

RQ2: How to mitigate the preference bias on LLMs?

- 이를 위한 Methodological Study를 진행함

  : ChatGPT and LLaMA2 70B

- Contact Hypothesis를 기반으로, LLM과 외부의 그룹 (external assistance)을 만나게 하였을 때 그 bias가 줄어들 것 이라는 가정
- 크게 두 카테고리의 메서드 사용

  : self-contact and external-contact

* 접촉 가설(Contact Hypothesis)
: 집단 간 편견을 감소시키기 위해 가장 포괄적으로 연구되어 온 대표적인 심리학 이론.
: 집단간의 접촉을 통해 집단간의 편견과 차별을 해소하고 관계를 개선할 수 있다는 이론

1) Self-contact approaches

: 3개의 method

1] Direct-Refine

: 모델한테 initial response를 주고 emotional support elements를 포함하게 refine하라고 instruct

2] Self-Refine

: feedback 기반으로 initial response를 refine

3] Emotional-CoT

: CoT처럼 현재 user의 감정 상태를 설명하고 response를 생성하게끔

RQ2: How to mitigate the preference bias on LLMs?

2) External-contact approaches

1] w/ COMET

external knowledge COMET (COMET-BART모델) 사용하여 외부 지식과 함께 생성하도록 instruct

2] w/ Example Expansion

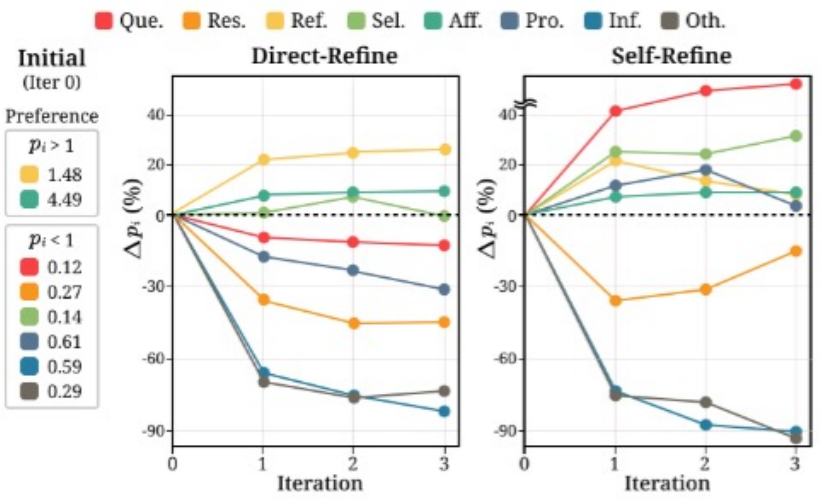
prompt에 2-shot examples → 4-shot examples

3] w/ Strategy Planner

fine-tune LLaMA2-7B as a strategy planner

RQ2: How to mitigate the preference bias on LLMs?

	Methods	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	B-2	R-L
	ChatGPT (*0-shot*)	13.50	1.38	6.27	14.86
Self	+ Direct-Refine	13.40	1.60	5.68	14.50
Self	+ Self-Refine	12.37	1.53	5.16	14.33
Self	+ Emotional-CoT	9.55	1.56	5.23	14.12
External	+ w/ COMET	12.78	0.95	6.71	15.07
External	+ w/ Example Expansion	16.91	0.82	**7.45**	**15.22**
External	+ w/ Strategy Planner	**21.09**	**0.36**	6.96	14.91
	LLaMA2-70B (*2-shot*)	14.55	0.47	6.15	14.29
Self	+ Direct-Refine	13.17	0.59	5.59	13.98
Self	+ Self-Refine	13.15	0.55	5.56	13.70
Self	+ Emotional-CoT	12.73	0.53	6.37	13.87
External	+ w/ COMET	14.53	0.51	6.21	14.55
External	+ w/ Example Expansion	15.14	0.44	**6.56**	**14.66**
External	+ w/ Strategy Planner	**21.09**	**0.36**	6.44	14.49



- self contact method 적용시

  : automatic metric에서 성능이 떨어짐

  : proficiency가 크게 떨어지고, preference bias가 더 심해지는 경향

  → 인간처럼, LLM이 bias를 가질 때 혼자 생각하는 것은 bias를 심화시킬 수 있음

- self contact method의 negative impact를 보기 위해서 iterative refinement

  : Direct-Refine과 Self-Refine에 대해서

  : iteration을 돌수록 initially preferred strategies에 대한 preference는 계속 증가, initially dispreferred strategies는 계속 감소

RQ2: How to mitigate the preference bias on LLMs?

	Methods	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	B-2	R-L
	ChatGPT (*0-shot*)	13.50	1.38	6.27	14.86
Self	+ Direct-Refine	13.40	1.60	5.68	14.50
Self	+ Self-Refine	12.37	1.53	5.16	14.33
Self	+ Emotional-CoT	9.55	1.56	5.23	14.12
External	+ w/ COMET	12.78	0.95	6.71	<u>15.07</u>
External	+ w/ Example Expansion	<u>16.91</u>	<u>0.82</u>	**7.45**	**15.22**
External	+ w/ Strategy Planner	**21.09**	**0.36**	<u>6.96</u>	14.91
	LLaMA2-70B (*2-shot*)	14.55	0.47	6.15	14.29
Self	+ Direct-Refine	13.17	0.59	5.59	13.98
Self	+ Self-Refine	13.15	0.55	5.56	13.70
Self	+ Emotional-CoT	12.73	0.53	6.37	13.87
External	+ w/ COMET	14.53	0.51	6.21	14.55
External	+ w/ Example Expansion	<u>15.14</u>	<u>0.44</u>	**6.56**	**14.66**
External	+ w/ Strategy Planner	**21.09**	**0.36**	<u>6.44</u>	<u>14.49</u>

Table 2: The results of methods on automatic metrics including $\mathcal{Q}$, $\mathcal{B}$, BLEU-2 (B-2) and ROUGE-L (R-L) for the entire test set (D). A single strategy planner is employed to predict strategies and provides them to each LLM. The best results of each LLMs are **bolded** and the second best are <u>underlined</u>.

- external contact method 적용시

  : closed, open source LLM에서 전반적으로 preference bias가 줄어든 경향

  : 하지만 COMET은 비교적 성능향상이 적거나 없음

  → Example Expansion, Strategy Planner와 같이 strategy에 대한 정보를 직접적으로 더 주는 것이 proficiency와 preference bias에 도움이 된다.

RQ2: How to mitigate the preference bias on LLMs?

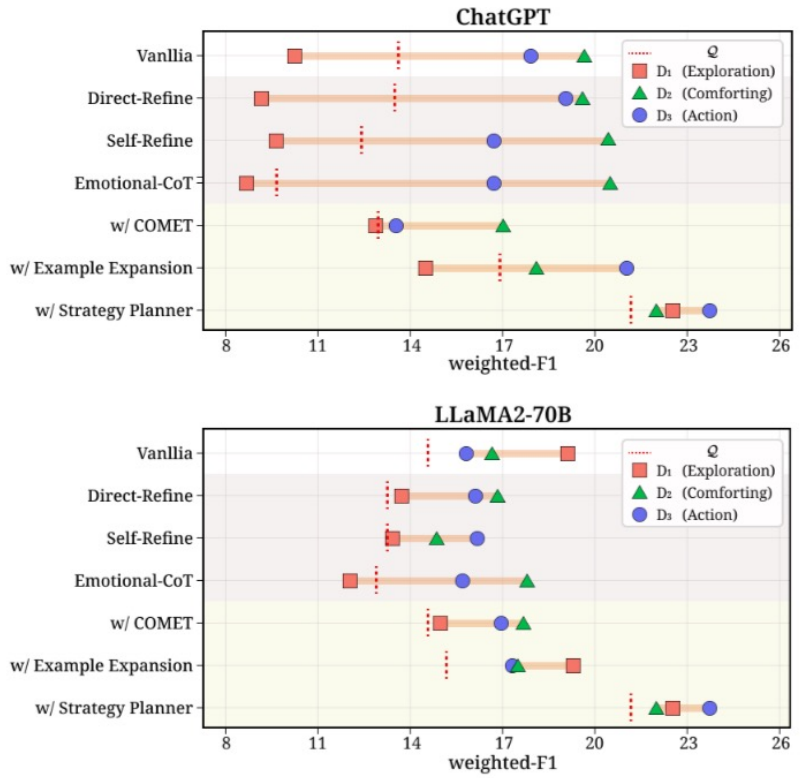


Figure 5: The weighted-F1 scores for each test set (D_t) and the macro-F1 score $\mathcal{Q}$ for the entire test set (D) on ChatGPT and LLaMA2. Self- and external-contact are backgrounded with gray and yellow, respectively.

- Each test set D_t에 대해서 self-contact(gray)와 external-contact(yellow) 성능

  : 이전 실험처럼 proficiency를 낮추고 preference bias를 높이던 self-contact 방법론에서 각 D_t 간의 차이가 크게 나타남

  → emotional support의 stage에 robust하지 못하기 때문에 poor-quality response를 생성할 수 있음

  : self-contact 시, D_1 에서 모든 방법론의 성능이 큰 차이로 떨어짐

  → Stage 구성이 Exploration → Comforting→ Action 이기 때문에 첫번째 stage에서 poor-quality response를 제공하게 되므로, 최종적으로 effective emotional support를 제공하기 어려울 수 있음

  : external-contact의 경우 각 stage간의 차이가 비교적 적음

  → 더 robust하고 effective emotional support가 가능함

RQ2: How to mitigate the preference bias on LLMs?

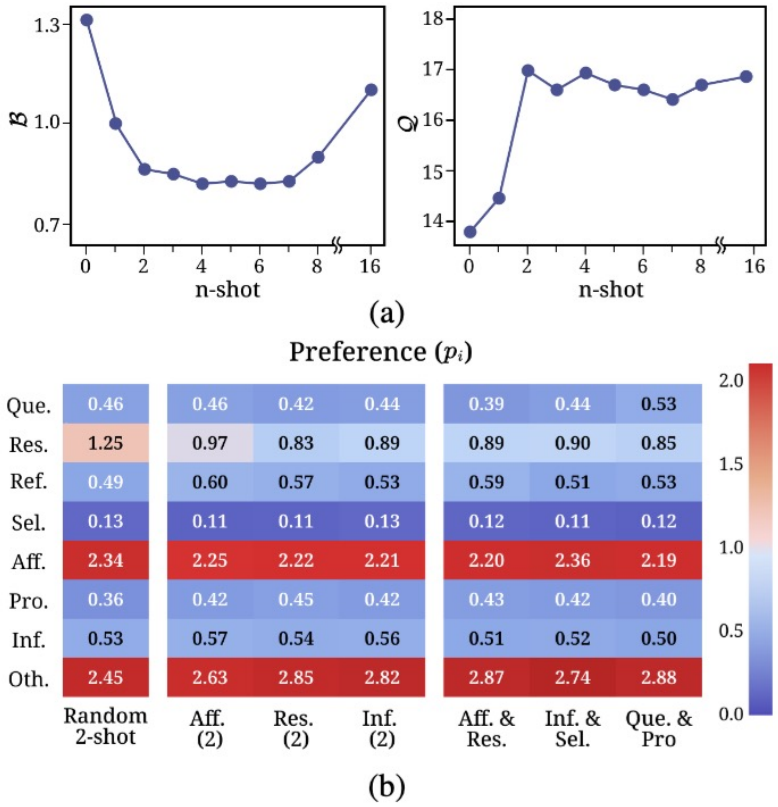


Figure 6: The results of (a) the variation in the number of shot examples, and (b) the effect of various combinations of strategies in 2-shot examples with ChatGPT.

- Effect of examples in the prompt

  : example 개수가 증가할수록 bias가 개선되고 성능이 향상되는 경향을 보이지만 8개 이상이 되면 bias가 크게 증가함

- Example의 strategy type에 따른 성능변화

  : various combinations of strategies within 2-shot example

  : 어떻게 sample을 구성하든 특정 strategy에 대한 편향은 존재하며, prompt 내의 sample에 큰 영향을 받지는 않음

RQ3: Does improving preference bias help to become a better emotional supporter?

- 정말 preference bias를 줄이는 게 도움이 되는가?
- Criteria of human evaluation

  : 심리학과랑 같이 helpful emotional support를 제공하고 있는 response인지 판단할 수 있는 criteria를 만듦

  emotional support의 목적 = user's state를 잘 파악하여 emotional intensity를 줄이는 것

  (1) **Acceptance**: Does the seeker accept without discomfort

  (2) **Effectiveness**: Is it helpful in shifting negative emotions or attitudes towards a positive direction

  (3) **Sensitivity**: Does it take into consideration the general state of the seeker.

  (4) **Alignment**: strategy와 response간의 alignment

RQ3: Does improving preference bias help to become a better emotional supporter?

ChatGPT	Acc.	Eff.	Sen.	Sat.
Vanilla	27.9	23.5	22.1	24.5
Tie	20.6	32.4	22.1	25.0
+ Self-Refine	**51.5**‡	**44.1**‡	**55.9**‡	**50.5**‡
Vanilla	22.9	24.0	14.6	20.5
Tie	21.9	33.3	27.1	27.4
+ w/ COMET	**55.2**‡	**42.7**†	**58.3**‡	**52.1**‡
Vanilla	13.1	25.3	16.2	18.2
Tie	26.3	26.3	21.2	24.6
+ w/ Example Expansion	**60.6**‡	**48.5**†	**62.6**‡	**57.2**‡
Vanilla	16.7	29.2	29.2	25.0
Tie	12.5	16.7	12.5	13.9
+ w/ Strategy Planner	**70.8**‡	**54.2**‡	**58.3**‡	**61.1**‡

Table 4: The results of comparative human evaluation between various methods applied to ChatGPT and vanilla ChatGPT. (†/‡: p-value < 0.1/0.05)

Methods	< 3 (*fail*)	≧ 3 (*acceptable*)
ChatGPT	**16.7**	83.3
+ Direct-Refine	**21.2**	78.8
+ Self-Refine	**17.4**	82.6
+ w/ Strategy planner	**8.0**	92.0
+ Oracle Strategy	**3.8**	96.2

Table 5: The ratio (%) of scores below 3 (fail) and scores of 3 or above (acceptable) in Seeker's Satisfaction (*Sat.*).

- 3명의 annotator한테 Win/Tie/Lose + 또 다른 3명한테 각 annotation이 적절한지 아닌지 평가하도록
- 기존 실험과 동일한 결과

  : 각 criteria의 평균인 Sat (seeker's satisfaction)를 기준으로 external-contact 가 self-refine보다 좋은 성능
- Self-Refine vs. w/ COMET

  : 기존 실험에서 proficiency는 유사하지만 preference bias에서 큰 차이로 w/COMET이 우수했음

  → human eval에서도 w/ COMET이 우수함
- Proficiency가 가장 우수하고 Preference bias가 가장 낮았던 w/ Strategy Planner

  → human eval에서도 가장 우수한 결과

→ Preference bias를 완화하는 것이 최종적으로 emotional support conversation에 중요하다

	Methods	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	B-2	R-L
	ChatGPT (*0-shot*)	13.50	1.38	6.27	14.86
Self	+ Direct-Refine	13.40	1.60	5.68	14.50
Self	+ Self-Refine	12.37	1.53	5.16	14.33
Self	+ Emotional-CoT	9.55	1.56	5.23	14.12
External	+ w/ COMET	12.78	0.95	6.71	15.07
External	+ w/ Example Expansion	16.91	0.82	**7.45**	**15.22**
External	+ w/ Strategy Planner	**21.09**	**0.36**	6.96	14.91

Conclusions

- strategy-centric analysis를 통해서 왜 LLM이 emotional support를 제공하는데 어려움을 겪는지, strategy의 중요성을 강조함

- LLM이 특정 strategy에 preference bias를 가지는 것을 증명
- LLM이 인간처럼 contact hypothesis 개념에 align되는 것을 보이고 external assistance가 preference bias를 완화할 수 있음을 보임
- Preference bias가 완화되면 appropriate strategy를 select하는데 robust해지기 때문에,
  emotional support의 quality가 향상되고 poor-quality response의 비율이 줄어듦을 강조함

FEEL: A Framework for Evaluating Emotional Support Capability with Large Language Models

Huaiwen Zhang*, Yu Chen*, Ming Wang and Shi Feng(✉)

School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
fengshi@cse.neu.edu.cn

Abstract

- Emotional Support Conversation (ESC)
  : effectively assist the user in mitigating emotional pressures
- 하지만 현재 inherent subjectivity때문에 emotional support capability를 측정하는 데 한계가 있다
  : human judgment와 low correlation

- emotional support capabilities를 측정하는 FEEL (Framework for Evaluating Emotional Support Capability with Large Language Models) 제안
  : ESC의 다양한 evaluation aspects를 고려하여 평가
  : probability distribution approach를 사용함으로써 안정적인 결과
  : 여러 LLM에 weight를 줘는 앙상블 전략을 통해 evaluation accuracy를 높임

- 기존의 ESC 모델의 발화들에 적용함으로써 FEEL의 성능을 평가
- 결과적으로 baseline과 비교하였을 때, FEEL에서 human evaluation과 alignment가 향상되었음

Introduction

- ESC가 mental health support, customer service chats에서 활용되고 있지만 evaluation은 여전히 challenging task로 여겨지고 있음

  : unreliable ESC evaluation systems은 사용자의 psychological stress를 증가시킬 수 있음

- 기존의 ESC evaluation은 크게 두가지

  1) traditional automatic metrics – BLEU, ROUGE, Distinct-n, METEOR

     : 복잡하고 다양한 인간의 감정을 이해, 평가할 수 없음

  2) training individuals to assess dialogue quality in specific aspects

     : 정해진 evaluation framework가 없으며, 주관적인 경험에 의존

- 최근 LLM을 기반으로 NLG evaluation method들이 제안됨

  : stochastic nature of computations 때문에 불안정한 경향

  : 대부분 일반적인 NLG evaluation, ESC를 위한 specialized prompt는 존재하지 않음

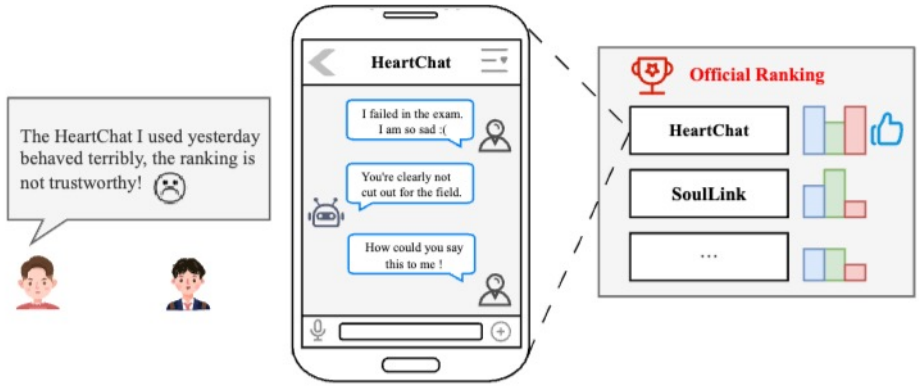


Fig. 1. Low-quality evaluation systems lead to poor emotional support experiences and even worsen user's situation.

Introduction

- Emotional support capability를 체계적으로 평가하기 위해 psychotherapy talk의 interaction을 분석하여 6가지 evaluation aspects 재정의

  : emotional support skills and text quality 측면으로

- a novel evaluation framework FEEL 제안

  : ERNIE-Bot 4.0, GLM-4, and GPT-3.5-Turbo – 3가지 모델을 모두 사용

  : LLM한테 task definition과 scoring criteria를 주고 평가하도록

  : score distribution을 여러번 뽑고 variance의 영향을 줄이기 위해 평균 -> 각 LLM의 평가 점수

  FEEL의 output = the score of emotional support capabilities

- Emotional support capability score dataset ESCEval 데이터셋 제안

Methodology – Task Definition

- categorize the evaluation of ESC into two distinct dimensions

  : emotional support skills and text quality

- Psychotherapy 개념을 기반으로 각 aspect를 구체화

  | emotional support skill

  1) Informativeness : supporter가 help-seeker에게 emotional problem을 자세히 설명하도록 하는 정도

  2) Comprehensibility : supporter가 help-seeker의 experiences와 feeling을 이해하는 정도

  3) Helpfulness: help-seeker의 emotional distress를 완화하고 조언을 할 수 있는 정도

  | Text Quality

  Consistency, Coherence, Safety

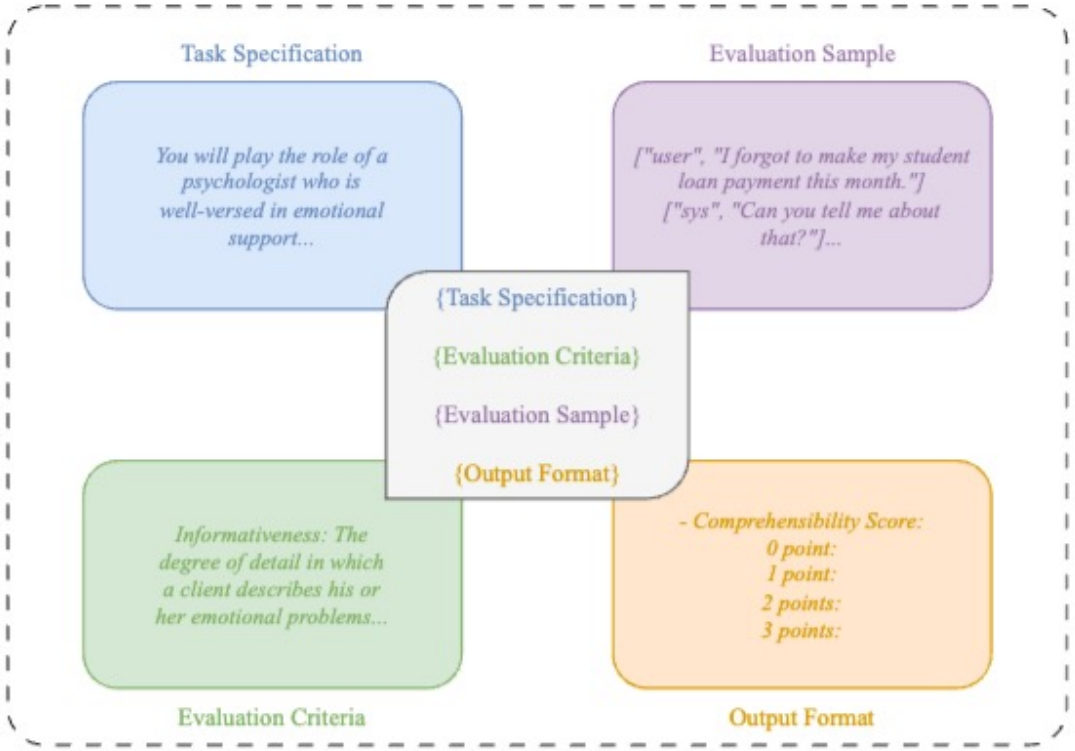
- 각 aspect에 대해서 four-tiered Likert scale score S_i, $S_i \in [0, 3]$

Methodology – Proposed FEEL

- FEEL: LLM을 기반으로 하는 comprehensive, multifaceted evaluator
- 총 3가지로 구성됨

  a) a structured prompt delineating task specification

  : task specification, 6가지의 aspect에 대한 criteria, evaluation sample dialogue, output format



Methodology – Proposed FEEL

- FEEL: LLM을 기반으로 하는 comprehensive, multifaceted evaluator
- 총 3가지로 구성됨

b) an advanced scoring algorithm

: 특정 scoring bands에 할당된 probabilities를 계산함으로써 각 LLM의 점수를 계산하는 scoring algorithm

: LLM의 답변 형태 – 총 합이 1이 되는 각 score에 대한 확률값

Answer format (give the probability of each score band for each type of score):
- Comprehensibility Score:
0 points:
1 point:
2 points:
3 points:

: 각 확률을 weight값으로 사용하여 score에 곱한 후 모두 더하여 해당 aspect에 대한 점수를 얻음

: 위 round를 총 10번 진행 후 평균값을 사용

→ LLM response의 변동성을 고려하는 방식

$$S_i = \frac{\sum_{n=1}^{10}\sum_{j=0}^{3} W_{j,n} * j}{10}$$

$W_{j,n}$: selection probability of score band j in iteration n

Methodology – Proposed FEEL

- FEEL: LLM을 기반으로 하는 comprehensive, multifaceted evaluator
- 총 3가지로 구성됨

c) a comprehensive integrative weighted computational approach

: 여러 LLM의 결과를 종합하여 최종 evaluation 점수를 계산

: preliminary testing을 했을 때 서로 다른 dialogue에서 모델별로 distinct advantage를 보였음

→ 각 모델의 점수에 weight를 주어 계산

: 각 LLM의 weight를 결정하기 위해, 각 LLM을 사용해서 ESCEval을 평가

: ESCEval 평가 결과와의 Spearman's rank correlation coefficient를 weight로 사용

$$\rho_{i,n} = \frac{c_{i,n}}{\sum_{n=1}^{3} c_{i,n}}$$

$$F_i = \sum_{n=1}^{3} \rho_{i,n} * S_{i,n}$$

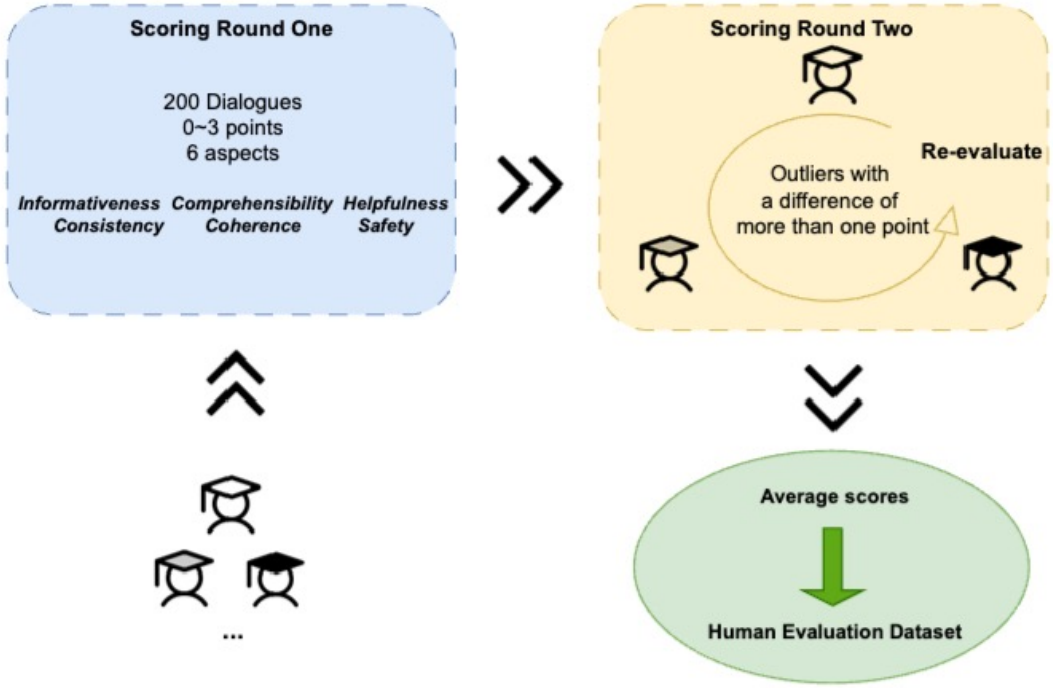
Methodology – ESCEval: A Dataset of Human ESC Evaluation

- 각 LLM의 emotional supporter capabilities를 평가하고 그 점수의 weight를 구하기 위해서는 reference standard가 필요
- ESCEval 데이터셋 구축

  ESConV, AUGESC에서 랜덤하게 뽑은 200개의 대화에 대해 평가

  총 6개의 aspect에 대해서 평가 후 평균 점수 계산

  two scoring rounds를 통해 주관성이 scoring에 미치는 영향을 최소화함



Methodology – Proposed FEEL

- FEEL: LLM을 기반으로 하는 comprehensive, multifaceted evaluator
- 총 3가지로 구성됨

c) a comprehensive integrative weighted computational approach

: 여러 LLM의 결과를 종합하여 최종 evaluation 점수를 계산

: preliminary testing을 했을 때 서로 다른 dialogue에서 모델별로 distinct advantage를 보였음

→ 각 모델의 점수에 weight를 주어 계산

: 각 LLM의 weight를 결정하기 위해, 각 LLM을 사용해서 ESCEval을 평가

: ESCEval 평가 결과와의 Spearman's rank correlation coefficient를 weight로 사용

Experiments and Results

- **Weight Determination and Analysis**

  : we employ the three LLMs' Spearman correlation coefficients and Kendall's tau coefficient on ESCEval to derive the final FEEL score

Table 1. Spearman's rank correlation coefficient (Spear.) and Kendall's tau coefficient (Kend.) for different models on the aspects of emotional support skills.

	Informativeness		Comprehensibility		Helpfulness	
	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.
ERNIE-4.0	0.368	0.270	0.372	0.269	0.414	0.313
GPT-3.5	0.163	0.119	0.190	0.139	0.360	0.264
GLM-4	0.364	0.299	0.317	0.250	0.385	0.297
FEEL	**0.404**	**0.300**	**0.429**	**0.314**	**0.509**	**0.377**

Table 2. Spearman's rank correlation coefficient (Spear.) and Kendall's tau coefficient (Kend.) for different models on the aspects of text quality.

	Consistency		Coherence		Safety	
	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.
ERNIE-4.0	0.427	0.323	**0.343**	**0.250**	0.384	0.298
GPT-3.5	0.126	0.088	0.135	0.094	0.257	0.192
GLM-4	0.313	0.244	0.265	0.211	0.311	0.252
FEEL	**0.434**	**0.327**	0.331	0.241	**0.409**	**0.314**

- 대부분의 correlation coefficients value가 0.3을 넘고 0.4 이상인 경우도 있음

  → 세 LLM이 다양한 측면에서 human evaluation과 높은 상관관계를 가짐
- FEEL이 weighted sum을 통해 각 모델의 장점을 종합하여 계산하기 때문에 가장 좋은 성능을 보임

- 하지만 특정 aspect (coherence)에서 데이터셋의 특성이나 모델 자체의 잠재적인 bias때문에 효과가 더 떨어질 수 있음

Experiments and Results

- **Comparative Results**

  : 기존 모델들의 output에 대한 평가 결과를 automatic metric과 비교

  : three pre-trained models - BlenderBot-Joint , MISC , **TransESC**

  : four large language models - Spark-V3.0, Baichuan2-Turbo, qwen-turbo, ChatGLM-6B

  : human evaluation – 각 aspect에서 각 모델 순위

  (1)Fluency, (2) Identification, (3) Empathy, (4) Suggestion and (5) Security

  : Evaluation Strategy

  1) Spearman's rank correlation coefficient
  2) Kendall's tau coefficient
  3) Root mean squared error
  4) Mean absolute error

model predictions(p_i)과 human rankings(r_i) 간의 차이를 측정하기 위해

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(p_i - r_i)^2}}{n}$$

$$MAE = \frac{\sum_{i=1}^{n}|p_i - r_i|}{n}$$

Experiments and Results

- Comparative Results

Table 3. The average Spearman's rank correlation coefficient (Spear.), Kendall's Tau (Kend.), Rooted Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on the sample.

	Spear.	Kend.	RMSE	MAE
BLEU-1	-0.136	-0.124	2.868	2.400
BLEU-2	-0.082	-0.076	2.878	2.343
ROUGE-1	-0.261	-0.210	3.145	2.714
ROUGE-2	-0.332	-0.257	3.230	2.743
ROUGE-L	-0.196	-0.162	3.017	2.543
METEOR	-0.029	-0.038	2.828	2.400
FEEL	**0.404**	**0.314**	**2.049**	**1.657**

→ FEEL is significantly better than all other baselines in four metrics.

Experiments and Results

- Ablation Experiment

Table 4. Ablation study results of FEEL.

	Spear.	Kend.	RMSE	MAE
ERNIE	0.219	0.187	2.324	1.714
GLM	0.161	0.124	2.505	2.086
GPT	0.182	0.174	2.126	1.900
ERNIE+GLM	0.247	0.200	2.342	1.886
ERNIE+GPT	0.264	0.181	2.331	1.857
GLM+GPT	0.386	0.276	2.150	**1.629**
FEEL	**0.404**	**0.314**	**2.049**	1.657

→ single-LLM, the combination of the two models (+FEEL) 과 비교

→ 두개의 모델만 사용하여 FEEL을 적용하여도 human 평가와 유사한 성능을 보임

Conclusion

- Emotional support capability를 측정할 수 있는 LLMs-based evaluator FEEL 제안
- ESC의 evaluation aspects를 정의하고 그에 따르는 high-quality human score dataset ESCEval 공개

- FEEL이 기존의 automatic evaluation metric보다 뛰어난 human alignment를 가지는 것을 증명
- Ablation experiment를 통해 여러 LLM을 사용하는 것이 evaluation quality를 높일 수 있음을 보임

Thank you