ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA241039

InsideOut: Unifying Emotional LLMs to Foster Empathy

Mikhail Mozikov[a,g,*], Nikita Severin[b,f,], Maria Glushanina[c], Mikhail Baklashkin[d], Andrey Savchenko[e,f] and Ilya Makarov[a,f,***]**

[a]AIRI, Moscow, Russia
[b]HSE University, Moscow, Russia
[c]SPbU, Saint-Petersburg, Russia
[d]MIPT, Moscow, Russia
[e]Sber AI Lab, Moscow, Russia
[f]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
[g]NUST MISIS, Moscow, Russia
ORCID (Andrey Savchenko): https://orcid.org/0000-0001-6196-0564, ORCID (Ilya Makarov): https://orcid.org/https://orcid.org/0000-0002-3308-8825

Abstract. This paper introduces InsideOut, an original innovative framework that augments the emotional intelligence of Large Language Models (LLMs). Motivated by the cartoon, InsideOut is designed around a net of specialized agents, each dedicated to one of Ekman's fundamental emotions. These agents collaboratively refine responses sensitive to the emotional context of interactions. Our assessments, conducted using EmpatheticDialogues and involving models like GPT-4 and GigaChat, indicate substantial improvements in identifying human emotions and generating empathetic responses. These improvements are most evident in situations with apparent valence-arousal differences. InsideOut offers a promising avenue for evolving AI into more perceptive and human-centric communicators.

1 Introduction

Integrating emotional intelligence into AI, particularly Large Language Models (LLMs), is crucial for human-AI interaction [34, 19]. LLMs are increasingly used in sectors like mental health support [9], customer service [23], and education [6]. However, studies suggest that mainstream LLMs, like ChatGPT by OpenAI, lack emotional intelligence compared to traditional models, particularly in generating empathetic dialogues [34, 19, 8].

We propose InsideOut, a system enhancing LLM-human interactions to tackle this challenge. It comprises emotionally enhanced LLM agents representing Ekman's five basic emotions [4]: anger, disgust, fear, happiness, and sadness. These agents collaborate to support a final aggregator, which formulates responses to users, enabling the system to deliver empathetic and relevant responses. We examine two configurations: Emotion Recognition in Conversation (ERC), where agents provide emotion labels and confidence levels, and Empathetic Response Generation (ERG), where variations of replies are generated. This setup separates emotion recognition [30, 31] and response generation, allowing users to prioritize and evaluate enhancements for ERC and ERG tasks separately.

* Corresponding Author Email: mb.mozikov@gmail.com
** Corresponding Author Email: nseverin14@gmail.com
*** Corresponding Author Email: iam.dadii.dh@gmail.com

InsideOut enhances emotion recognition and empathy in LLMs, integrating easily with existing systems. Rigorously tested against baseline LLM adaptations and specialized ERC models, it shows transformative potential in advancing emotionally intelligent AI. It offers a straightforward upgrade path for individuals and businesses, promising more authentic connections with human emotions.

Thus, our main contributions are as follows:

- Introduction of InsideOut, a system that enhances the emotional intelligence of LLMs, allowing for more empathetic interactions.
- Development of a framework that supports easy experimentation and comparison of various InsideOut-enhanced LLM configurations, thus simplifying access to sophisticated emotional AI tools.
- Demonstration of significant improvements in emotion recognition and empathetic response generation, with state-of-the-art accuracy surpassing existing models.

2 Related Work

Emotion Recognition in Conversation (ERC). Research in ERC has made significant strides, especially with the most recent LLM developments. The authors of [3] initially proposed M2FNet, a multimodal fusion network for ERC, incorporating visual and audio data alongside textual information [29, 30]. Building on that, [35] introduced DialogueLLM, which fine-tunes LLMs with multimodal emotional dialogues, enhancing emotion understanding with contextual and visual cues. InstructERC [10] also used LLMs to introduce a retrieval template module that integrates multi-granularity dialogue supervision and unifies emotion labels via a feeling wheel. These works underscore the evolution of ERC towards more comprehensive and context-aware systems.

Empathetic Response Generation (ERG). The empathetic response generation task involves interpreting user emotions and responding with empathy [27]. Previous work has focused on emotional state identification [27], strategic response generation [17], positive emotion elicitation [38], and integrating contextual knowledge [14]. Recent advances include emotion-cause elements [2], refining responses by desired emotions [26], and developing emotion-

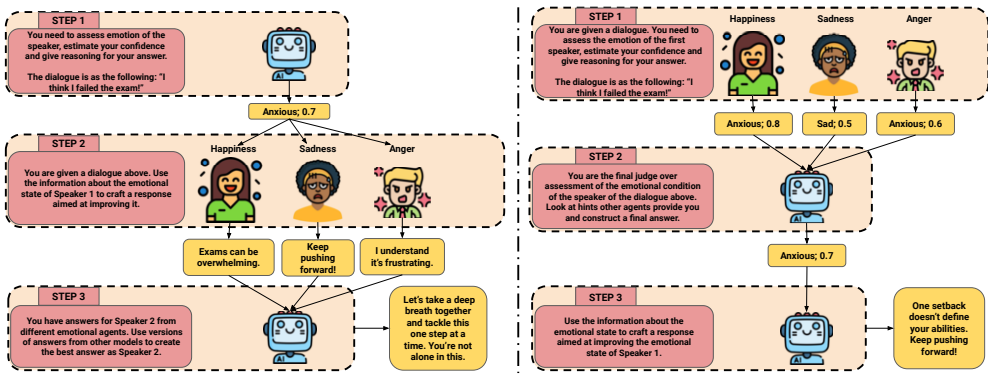

Figure 1: InsideOut for ERC (left) and ERG (right) tasks. Each emotional agent is tasked to embody a specific emotion, guiding its decisions accordingly. In the ERC task, agents are prompted to identify human emotions from the conversation history. In the ERG task, agents are tasked with predicting the final response in the conversation based on the current conversation history and extracting other speaker's emotions.

ally self-aware models [36]. In contrast with previous research, our framework employs a network of emotionally adjusted agents for more accurate and empathetic responses.

3 Proposed Framework

This section presents the proposed InsideOut plug-and-play framework (Fig. 1), which can extend any existing LLM. It comprises two distinct configurations tailored to address specific tasks in emotional AI: ERC and ERG. The core architecture includes emotional agents and an Aggregate Agent organized in a star-like network. Each emotional agent is prompted [11, 20] to make decisions based on one of five basic emotions (anger, sadness, happiness, disgust, and fear) selected according to Paul Ekman's established classification [5]. These agents analyze various emotional contexts from different perspectives, while the Aggregate Agent synthesizes these perspectives based on the task requirements.

In the ERC configuration, each emotional agent is prompted with dialogue and tasked with determining the underlying emotion. They provide evaluations of the emotion or a probabilistic confidence level and a rationale for their assessment. The Aggregate Agent consolidates these outputs to form a comprehensive judgment on the user's emotional state, which the Simple Agent then utilizes. This agent, equipped with emotional context and dialogue history, crafts responses designed to engage the user positively. This configuration is optimized for robust performance in emotion recognition tasks.

In the ERG setting, the LLM assesses the user's emotional state in a zero-shot manner and shares this with emotional agents. These agents, informed by the assumed emotions, propose responses aimed at improving the user's well-being. The Aggregate Agent then selects the most effective response, optimizing for empathy.

The proposed methodology differentiates between emotion recognition and positive emotion elicitation, acknowledging their unique influences on conversational AI dynamics [38]. By individually optimizing these aspects, we can examine their effects on system performance, enhancing emotional intelligence and user engagement.

4 Experimental setup

Dataset. Our experiments are conducted on the EmpatheticDialogues dataset [27], a comprehensive collection of multi-turn conversations comprising 25000 empathetic dialogues. These dialogues were gathered through Amazon Mechanical Turk and involved interactions between a speaker and a listener. The dataset includes 32 emotion labels evenly distributed capturing speakers' and listeners' emotional states. In the same paper, human assessors estimated models trained on this dataset to be more empathetic.

Selected LLMs. We integrate our framework into different classes of LLMs, including GPT-3.5 and GPT-4 by OpenAI, which are known for their power and wide usage. We expect the best results given GPT-4's superior performance in various tasks [25, 22, 7]. Additionally, we explore Mistral-7B, a smaller model showing improvements over larger ones, for the ERC task. We also tested sensitivity to language distribution using GigaChat and trained on more Russian texts for Russian language processing. For the sake of reproducibility, in all our experiments, we fixed the versions of the proprietary models ("gpt-3.5-turbo-0125" for GPT-3.5 and "gpt-4-0125-preview" for GPT-4) and set the temperature parameter to 0.

Task-specific baselines. Following prior literature [1, 32], we use various state-of-the-art baselines to measure our model's performance. **Transformer** [33] is a basic Transformer-based encoder-decoder model. **MoEL**[16] detects user emotions and generates an empathetic response by combining output states from specialized Listeners optimized for different emotions. **MIME**[18] model incorporates polarity-based emotion clusters and stochastic emotion mixture. **EmpDG** [13] leverages coarse-grained dialogue and token-level emotions. **CEM** [28] includes the user's emotional and cognitive understanding of their situation in the model configuration. **KEMP** [14] uses the CONCEPT-NET knowledge graph and the emotional lexicon "NRC VADas" to enhance implicit emotion representation.

Evaluation: Classic metrics. Following the literature [1], we provide the accuracy measure (ACC) for the ERC task and use variations of the ROUGE [15], BLEU [24], and Distinct-1 [12] metrics for the ERG task. ROUGE and BLEU are commonly used to measure similarity of the generated response to the reference text, while the Distinct-1 metric measures the generation diversity.

Evaluation: Assessor Metric. While classical ERG metrics focus on n-gram matching and lack deeper semantic understanding, human assessors offer holistic evaluations, considering context and novelty. However, human evaluation is costly and impractical for large datasets. Recent studies employ advanced language models like GPT-4 for automated evaluation [21], which is cost-effective and scalable. According to [37], evaluation criteria include Fluency (response coherence and smoothness), Identification (effectiveness in addressing the seeker's problems), Empathy (understanding of the seeker's feelings and situation), Suggestion (quality of advice given), and Overall effectiveness in providing emotional support.

5 Experimental Results

Let us present our experiments' detailed results, which aim to evaluate the influence of the framework. We will refer to "Baseline" to denote the configuration involving the LLM alone and "InsideOut" to describe the LLM enhanced by our framework. In the ERG task, we also employ the notation "ERC+Baseline/InsideOut" to refer to the configuration where emotions are initially recognized and incorporated as supplementary input to the LLMs alongside the dialogue history. The results for each task are presented separately.

Table 1: Result of automatic evaluation of Baseline LLMs in ERC task (32 classes), ACC. Both LLMs significantly outperform conventional SOTA baselines.

MoEL	MIME	EmpDG	CEM	KEMP	Baseline GPT-3.5	Baseline GPT-4
31.74	30.96	31.65	36.84	36.57	38.00	**44.20**

Table 2: Result of automatic evaluation of Baseline LLMs and its modifications in ERC task, ACC. Consistent improvement for GPT-4 and GigaChat models with InsideOut modification is observed.

# Classes	Base Model	Baseline	InsideOut	Improv.
32 classes	GPT-3.5	38.00	35.30	-7.11%
	GPT-4	44.20	45.10	2.04%
	GigaChat	30.79	33.31	**8.18%**
	Mistral-7B	18.85	16.42	-12.89%
18 classes	GPT-3.5	41.70	50.04	**20%**
	GPT-4	51.72	55.2	6.73%
	GigaChat	41.24	46.35	12.39%
	Mistral-7B	31.15	30.71	-1.41%

Emotion Recognition in Conversations (ERC). Table 1 and 2 show the results for the ERC task. Here and further, the best results are highlighted in bold, and the second-place results are underlined.

A study comparing GPT-3.5, GPT-4, and state-of-the-art models on the EmpatheticDialogues dataset shows LLMs to outperform significantly, with GPT-4 leading by 21% in emotion prediction. InsideOut consistently benefits GPT-4 and GigaChat across 32 emotion classes. GPT-3.5 improves in an 18-emotion subset but struggles in the 32-class setting. Mistral-7B's performance decreases with task complexity, while GPT-4 handles it effectively.

Detailed analysis of responses highlights a challenge with baseline models: they tend to misclassify emotions as happiness, particularly overlooking anger and sadness, consistent with Ekman's theory [4]. Baseline GPT models struggle with nuanced emotional distinctions. InsideOut framework addresses this problem, improving accuracy across all emotions, including finer classifications.

Table 3: Result of automatic evaluation with classic metrics in ERG task: BLEU (B-1, B-2, B-3, B-4), ROUGE (R-1, R-2) and Distinct-1 (Dist-1). While LLMs score lower on text similarity metrics, higher Dist-1 suggests greater answer variability.

Base Model	Method	B-1	B-2	B-3	B-4	R-1	R-2	Dist-1
-	Transformer	18.07	8.34	4.57	**2.86**	17.22	4.21	0.36
	MoEL	18.07	8.30	4.37	2.65	**18.24**	**4.81**	0.59
	MIME	18.60	8.39	4.54	2.81	17.08	4.05	0.47
	EmpDG	**19.96**	**9.11**	**4.74**	2.80	18.02	4.43	0.46
	CEM	16.12	7.29	4.06	2.03	15.77	4.50	0.62
	KEMP	16.72	7.17	3.77	2.33	16.11	3.31	0.66
GPT-3.5	Baseline	11.11	2.77	0.83	0.29	16.29	2.25	0.84
	ERC+Baseline	11.84	2.85	0.84	0.3	16.91	2.3	**0.86**
	ERC+InsideOut	8.74	1.89	0.46	0.14	14.02	1.56	0.78
GPT-4	Baseline	7.27	1.6	0.41	0.15	11.08	1.34	0.83
	ERC+Baseline	7.69	1.6	0.41	0.11	11.57	1.3	0.83
	ERC+InsideOut	6.11	1.26	0.24	0.06	10.16	1.01	0.74

Empathetic Response Generation (ERG). The results for the ERG task are outlined in Tables 3 and 4 that provide insights from classic metrics and GPT-4's assessment.

Table 4: ERG Evaluation Metrics by GPT-4 as an Assessor, score out of 10. Columns: Fluency (F), Identification (I), Empathy (E), Suggestion (S), Overall (Ovr.) and Improvement Overall (Impr. Ovr.).

Model	Method	F	I	E	S	Ovr.	Impr. Ovr.
GPT-3.5	Baseline	9.19	8.02	8.98	6.02	8.06	-
	ERC+Baseline	9.21	7.95	8.84	5.88	7.96	-1.24%
	ERC+InsideOut	**9.22**	**8.05**	**9.05**	**7.11**	**8.28**	2.73%
GPT-4	Baseline	9.16	8.04	8.96	6.72	8.14	-
	ERC+Baseline	9.14	8.06	9.00	6.72	8.19	0.61%
	ERC+InsideOut	**9.24**	**8.10**	**9.24**	**7.53**	**8.46**	3.93%

One notable observation is the inconsistency in various metrics. Traditional metrics like ROUGE and BLEU suggest that task-specific models outperform LLMs in the ERG task. In contrast, the Dist-1 metric indicates LLMs produce more varied responses than baseline models, aligning with prior research [37]. Exploring InsideOut's impact on LLMs, Table 4 shows it enhances overall response quality by over 2.7%. Interestingly, integrating InsideOut reduces the Dist-1 metric, indicating more human-like diversity in responses. This highlights the trade-offs between traditional text similarity metrics and emotional resonance in generated responses, underscoring InsideOut's potential to enhance empathy in LLM outputs.

6 Conclusion

This study introduces InsideOut, a framework designed to increase LLMs' empathy. We comprehensively examine its impact on various LLMs for ERC and ERG tasks. Our findings show a complex landscape in which the benefits of employing InsideOut vary significantly across different models and tasks.

In ERC tasks, InsideOut significantly boosts emotion recognition accuracy in models like GPT-4 and GigaChat, particularly in distinguishing among 18 different emotion classes in varied emotional contexts, achieving up to 20% improvement over baseline models. However, challenges arise with models like Mistral-7B or when the emotional classes are closely clustered, showcasing difficulties in adapting InsideOut to diverse model capacities.

The ERG task yields mixed results, balancing text-based accuracy metrics and empathetic response quality. While BLEU and ROUGE typically decrease with InsideOut, there are notable improvements in empathetic and contextually relevant metrics. This highlights trade-offs between traditional text similarity metrics and emotional resonance in responses. Our study underscores InsideOut's potential to enhance LLMs' understanding and generation of emotionally resonant content, which is vital for empathetic interactions. Future work should refine integration techniques for smaller models like Mistral-7B to boost InsideOut's utility in practical applications.

Overall, this research contributes to a deeper understanding of how LLMs can be effectively augmented to better handle the complexities of human emotional expression, providing a stepping stone for future innovations in empathetic AI systems.

Limitations. Although our experiments used only the EmpatheticDialogues dataset, the results strongly support InsideOut's effectiveness. To further generalize these findings, future research should consider testing InsideOut on diverse datasets and with a broader scope of models, including small and mid-size LLMs.

Acknowledgements

We would like to express our sincere gratitude to Valeria Bodishtianu for her invaluable assistance with the preparation of this paper. We appreciate her significant contributions to writing the text, developing research ideas, and assisting with experimental design.

References

[1] H. Cai, X. Shen, Q. Xu, W. Shen, X. Wang, W. Ge, X. Zheng, and X. Xue. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics (ACL)*, pages 7858–7873, Toronto, Canada, 2023. Association for Computational Linguistics.

[2] Y. Chen and C. Liang. Wish I can feel what you feel: A neural approach for empathetic response generation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics (EMNLP)*, pages 922–933, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[3] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation, 2022.

[4] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6 (3-4):169–200, 1992.

[5] P. Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[6] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[7] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382 (2270):20230254, 2024.

[8] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023.

[9] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models, 2023.

[10] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework, 2024.

[11] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie. Large language models understand and can be enhanced by emotional stimuli, 2023.

[12] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models, 2016.

[13] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen. EmpDG: Multi-resolution interactive empathetic dialogue generation. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.394. URL https://aclanthology.org/2020.coling-main.394.

[14] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen. Knowledge bridging for empathetic dialogue generation, 2021.

[15] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[16] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung. Moel: Mixture of empathetic listeners, 2019.

[17] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang. Towards emotional support dialog systems, 2021.

[18] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria. MIME: MIMicking emotions for empathetic response generation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.721. URL https://aclanthology.org/2020.emnlp-main.721.

[19] S. Marcos-Pablos and F. J. García-Peñalvo. Emotional intelligence in robotics: A scoping review. In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection 1*, pages 66–75. Springer, 2022.

[20] M. Mozikov, N. Severin, V. Bodishtianu, M. Glushanina, M. Baklashkin, A. V. Savchenko, and I. Makarov. The good, the bad, and the hulk-like gpt: Analyzing emotional decisions of large language models in cooperation and bargaining games. *arXiv preprint arXiv:2406.03299*, 2024.

[21] B. Naismith, P. Mulcaire, and J. Burstein. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, 2023.

[22] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[23] K. Pandya and M. Holia. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations, 2023.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[25] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[26] Y. Qian, B. Wang, S. Ma, W. Bin, S. Zhang, D. Zhao, K. Huang, and Y. Hou. Think twice: A human-like two-stage conversational agent for emotional response generation, 2023.

[27] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: a new benchmark and dataset, 2019.

[28] S. Sabour, C. Zheng, and M. Huang. Cem: Commonsense-aware empathetic response generation, 2021.

[29] A. Savchenko, A. Alekseev, S. Kwon, E. Tutubalina, E. Myasnikov, and S. Nikolenko. Ad lingua: Text classification improves symbolism prediction in image advertisements. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1886–1892, 2020.

[30] A. V. Savchenko. EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5716–5724, 2023.

[31] V. Savchenko and A. Savchenko. Information-theoretic analysis of efficiency of the phonetic encoding–decoding method in automatic speech recognition. *Journal of Communications Technology and Electronics*, 61:430–435, 2016.

[32] L. Sun, N. Xu, J. Wei, B. Yu, L. Bu, and Y. Luo. Rational sensibility: Llm enhanced empathetic response generation guided by self-presentation theory, 2024.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[34] K. Yellapantula and M. Ayachit. Significance of emotional intelligence in the era of artificial intelligence: A study on the application of artificial intelligence in financial and educational services sector. *Ushus - Journal of Business Management*, 18:35–48, 01 2019. doi: 10.12725/ujbm.46.3.

[35] Y. Zhang, M. Wang, Y. Wu, P. Tiwari, Q. Li, B. Wang, and J. Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations, 2024.

[36] W. Zhao, Y. Zhao, X. Lu, and B. Qin. Don't lose yourself! empathetic response generation via explicit self-other awareness. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.843. URL https://aclanthology.org/2023.findings-acl.843.

[37] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.

[38] J. Zhou, Z. Chen, B. Wang, and M. Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, Toronto, Canada, 2023. Association for Computational Linguistics.