APPENDIX

A. Dataset Annotations distributions

Figure 5 and 4 demonstrate the normalized distribution of emotion groups and intents for speakers, listeners, and reflective responses.

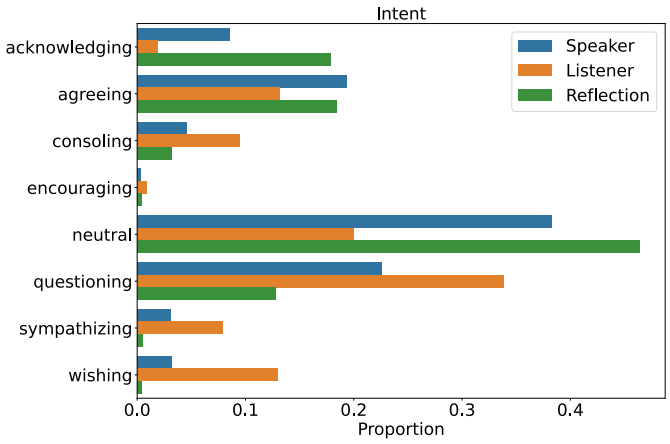

Fig. 4. Normalized Distribution of Intent for Speaker, Listener, and Reflection.

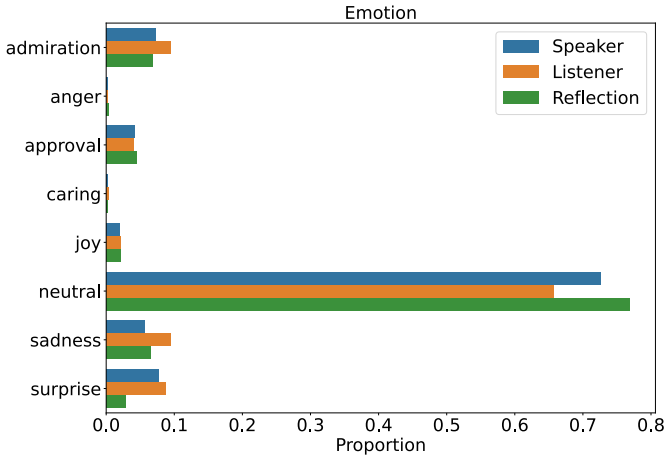

Fig. 5. Normalized Distribution of Emotion for Speaker, Listener, and Reflection.

B. Automated Evaluations Error Bars

The error bars for automated evaluation of Llama-2 based models are provided in Table III.

C. Human Evaluation for GPT-2 based Models

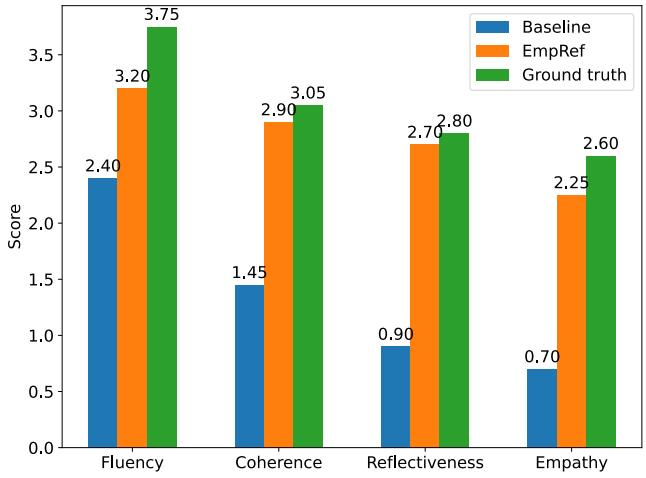

Fig. 6. Human evaluation results for GPT-2 based EmpRef and Fine-tuned models. The mean scores on human evaluation metrics.

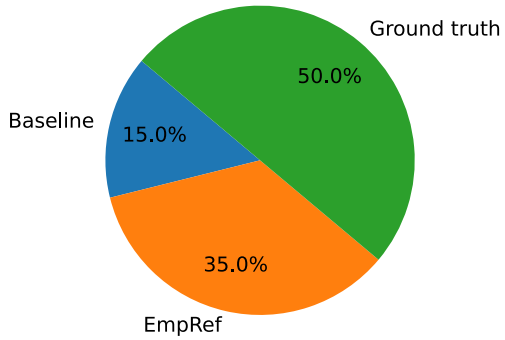

Fig. 7. Human evaluation results for GPT-2 based EmpRef and Fine-tuned models. The percentage being ranked as the best responses.

This section presents the results of human evaluations for GPT-2 based models. Figure 6 displays the average scores given to responses from the ground truth (target response in test dataset), EmpRef, and the fine-tuned baseline for each metric. The few-shot performance of GPT-2 is so unreliable that we excluded it from the human evaluation results, as nearly 20% of the responses were trapped in a repetitive loop. While the ground truth responses achieve the highest average scores across all metrics, the generations from our proposed model, EmpRef, closely follow and significantly outperform the baseline generations in all criteria. A noticeable gap is observed in the categories of Reflectiveness and Empathy, where EmpRef achieves scores three times better than the baseline. However, in comparison to other metrics, we notice a larger discrepancy between the fluency of EmpRef generations

TABLE III
AUTOMATED EVALUATION RESULTS FOR DIFFERENT METHODS WITH ERROR BARS. EACH VALUE REPRESENTS THE MEAN SCORE ± STANDARD DEVIATION.

Method	**BLEU-2**	**ROUGE-L**	**METEOR**	**Distinct-1**
Few-shot	0.0209 ± 0.0029	0.1123 ± 0.0062	0.1442 ± 0.0092	0.7770 ± 0.0062
Fine-tuned	0.0581 ± 0.0092	0.1580 ± 0.0128	0.2450 ± 0.0231	0.7072 ± 0.0100
Prepend	0.0882 ± 0.0101	0.1965 ± 0.0139	0.2118 ± 0.0175	0.6943 ± 0.0190
EmpRef	0.2408 ± 0.0141	0.3575 ± 0.0164	0.4798 ± 0.0197	0.7410 ± 0.0137

and the ground truth. We hypothesize that our model's focus on generating empathetic responses, guided by additional embeddings, may unintentionally compromise the fluency of the output by overshadowing the generation of grammatically correct and natural-sounding sentences. Figure 7, a pie chart illustrates the origin of the best responses, indicating that 50% of the highest-rated responses are from ground truth, 35% are from EmpRef, and only 15% are from fine-tuned baseline. Furthermore, the Cohen kappa level of agreement between annotators is substantial, with a value of 0.64.