Table 5: Label distribution of datasets. The labels in both datasets are biased towards neutral.

Dataset	DailyDialog	reconstructed MEmoR
Positive	12.7%	8.7%
Neutral	84.6%	80.3%
Negative	2.7%	11.0%
Total	100.0%	100.0%

Table 6: Train/Valid/Test split.

Dataset	**Task**	**Train**	**Valid**	**Test**
DailyDialog	EERC	85,570	7,962	6,632
	EEFC	74,548	6,973	6,632
reconstructed	IERC	4,767	585	573
MEmoR	IEFC	7,810	742	573

A Dataset Details

We show the label distribution of each dataset in Table 5 and the number of data for each task in Table 6. The datasets were split in the same way as the original data for both DailyDialog and reconstructed MEmoR. The train and validation data sizes for EEFC are smaller than those for EERC, and IERC than IEFC. This is because EEFC and IERC require two annotated utterances as the input (i.e., the current utterance and the next emotion, the current emotion and the next utterance). As for the test data, we used the same data for EERC and EEFC, and for IERC and IEFC to compare the results between these tasks.

B Hyperparameters

The hyperparameters are shown in Table 7. All the models were trained with one GPU (NVIDIA A100). At the end of the training of each task, we loaded the model of the epoch that achieved the highest macro-F1 score on the validation dataset. We fine-tuned Llama 2 using LoRA (Hu et al., 2021).

Table 7: Hyperparameters.

Model	Task	Input Variation	Learning Rate	Batch Size	Epoch
Llama-2-13b-hf	IEFC	full history	1e-5	4	10
		last uttr	2e-5	2	
		no history	2e-5	1	
DistilRoBERTa-base	EERC	all	warmup from 0 to 5e-05	64	40
	EEFC			64	60
	IERC			128	40
	IEFC			128	40