Qualitative Evaluation of Generated Responses

Setup and Evaluation Criteria

We compare two systems on emotionally aligned conversation generation:

i. *Corpus-based chatbot (Q1)*

Retrieves the most similar utterance from the training corpus using weighted similarity across:

- text embedding similarity
- emotion intensity
- empathy
- emotional polarity

ii. *ICL LLM chatbot (Q2)*

- Uses few-shot prompting with an instruction-tuned LLM.
- Prompt contains the first 5 turns + emotional targets + 3 example conversations.

Five random conversations from the development set were analyzed (turns 6–10).

Evaluated dimensions (1–5 scale):

- Fluency
- Relevance
- Coherence
- Emotional alignment

Conversation C-1 – Foster System / Shelton's Journey

Corpus-based model:

- Generic, repetitive questions ("How did you feel about the article?")
- Weak emotional expression
- Topic drift in later turns

ICL model:

- Minor prompt leakage on first turn
- Better alignment with empathy and emotional framing
- More coherent and topical

Ratings:

Fluency: Corpus 4 / ICL 4
Relevance: Corpus 2 / ICL 3
Coherence: Corpus 2 / ICL 3
Emotional alignment: Corpus 2 / ICL 3

Dimension	Corpus-based	ICL LLM
Fluency	4	4
Relevance	2	3
Coherence	2	3
Emotional alignment	2	3

Conversation C-2 – Celebrity Relationship / Insecurity

Corpus-based model:

- Off-topic drift
- Neutral emotional tone
- Limited empathy expression

ICL model:

- Fluent, coherent, and thematically appropriate
- Better modeling of insecurity and public pressure

Ratings:

Fluency: Corpus 4 / ICL 5
Relevance: Corpus 3 / ICL 4
Coherence: Corpus 3 / ICL 4
Emotional alignment: Corpus 3 / ICL 4

Dimension	Corpus-based	ICL LLM
Fluency	4	5
Relevance	3	4
Coherence	3	4
Emotional alignment	3	4

Conversation C-3 – Animal Cruelty (Flamingo Case)

Corpus-based model:

- Repetitive sadness statements
- Underestimates emotional intensity

ICL model:

- Strong negative polarity and high empathy
- More specific moral commentary

Ratings:

Fluency: Corpus 4 / ICL 4
Relevance: Corpus 3 / ICL 4
Coherence: Corpus 3 / ICL 4
Emotional alignment: Corpus 3 / ICL 4

Dimension	Corpus-based	ICL LLM
Fluency	4	4
Relevance	3	4
Coherence	3	4
Emotional alignment	3	4

Conversation C-4 – Environmental Disaster / Cleanup

Corpus-based model:

• Basic relevance but formulaic
• Moderate empathy expression

ICL model:

• Coherent and well-structured
• Occasionally too generic or global in perspective

Ratings:

Fluency: Corpus 4 / ICL 5
Relevance: Corpus 3 / ICL 3
Coherence: Corpus 3 / ICL 4
Emotional alignment: Corpus 3 / ICL 3

Dimension	Corpus-based	ICL LLM
Fluency	4	5
Relevance	3	3
Coherence	3	4
Emotional alignment	3	3

Conversation C-5 – Social / Moral Topic

Corpus-based model:

- High repetition
- Weak emotional nuance

ICL model:

- Better flow and emotional alignment
- Sometimes generic motivational tone

Ratings:

Fluency: Corpus 4 / ICL 5
Relevance: Corpus 3 / ICL 4
Coherence: Corpus 2 / ICL 4
Emotional alignment: Corpus 3 / ICL 4

Dimension	**Corpus-based**	**ICL LLM**
Fluency	4	5
Relevance	3	4
Coherence	2	4
Emotional alignment	3	4

Emotion / Empathy / Polarity Analysis

Corpus-based:

- Low-intensity, generic responses → classifier predicts weaker emotion/empathy.
- Frequent neutral polarity even when gold labels are strongly positive/negative.

ICL model:

- Uses emotionally rich language → closer alignment with gold polarity & intensity.
- Higher empathy due to explicit acknowledgment of suffering or hardship.

Overall Conclusion

The ICL LLM chatbot outperforms the corpus-based chatbot across all qualitative dimensions.

It provides:

- More human-like fluency
- Stronger coherence across turns 6–10
- Better emotional alignment
- Higher empathy expression

The corpus-based model is simple and stable but suffers from:

- Repetitive template-like generations
- Poor adaptation to emotional cues
- Weak long-range coherence

Therefore, the ICL LLM chatbot is the recommended approach for generating emotionally grounded conversations.

GitHub Link:
https://github.com/VishaalD07/Empathic-Dialogue-Generation-on-WASSA-2024-Track-2