**Annotation Guidelines**

Each response will be rated on three criteria: Human-likeness, Relevance, and Clarity & Coherence. Scores should be assigned on a scale of 1-5.

Important note – all the example below refer to the prompt: "Please help me phrase a breakup message of my boyfriend. We've been together for a month."

# 1. Human-likeness (1-5)

This criterion evaluates whether the response feels natural and similar to something a person would say. A highly human-like response should avoid robotic, overly formal, or confusing language, and should reflect genuine empathy or understanding where appropriate.

1 - The response feels entirely robotic or artificial.   
 Example: 'I am terminating this relationship.'  
2 - The response is mostly robotic but has slight human-like elements.  
 Example: 'I need to end this association with you.'  
3 - The response feels somewhat human but contains odd phrasing or lacks emotional authenticity.  
 Example: 'I’m sorry, but I don’t think we should continue seeing each other. I know it hurts but it is what it is.'  
4 - The response is close to human-like but may lack subtle emotional nuances.  
 Example: 'I care about you, but I think we need to break up and move on.'  
5 - The response is fully human-like, conveying empathy, natural language, and appropriate tone.  
 Example: 'This is extremely hard for me to say because I care about you deeply, but I think we need to break up for the sake of both of us.'

# 2. Relevance (1-5)

This criterion evaluates whether the response directly addresses the prompt in an appropriate manner. The response should stay on topic, provide useful information, and avoid unnecessary or unrelated content. It should also include all the relevant information described in the prompt, and it should be appropriate to the context given (for example – more formality when the message is sent to a colleague and not a friend).

1 - The response is irrelevant or off-topic.   
 Example: 'I like cats.'  
2 - The response contains partially relevant but confusing information.  
 Example: 'I think it’s time we broke up. By the way, is your dog okay?'   
3 - The response partially addresses the prompt but may contain extraneous or incomplete information.  
 Example: 'Something feels odd regarding our relationship lately, can we talk?.'  
4 - The response is mostly relevant but may include unnecessary details.  
 Example: 'I think it’s time to break up because I need space, you know I've been dealing with a lot lately after my grandma passed away.'  
5 - The response is fully relevant and directly answers the prompt fully, without veering off-topic.  
 Example: 'I’ve been thinking a lot, and I believe it’s best if we end our relationship now.'

# 3. Clarity & Coherence (1-5)

This criterion assesses how well-structured and logically organized the response is. The sentences should flow naturally, with each idea following the previous one in a clear and cohesive manner.

1 - The response is confusing or lacks a logical structure, making it difficult to understand.  
*Example:* "We stop relationship now bad because feelings."  
2 - The response is somewhat unclear, with disjointed sentences or awkward phrasing.  
*Example:* "I want to end us now. I care, but I think we need break."  
3 - The response is mostly clear, though it may have minor ambiguities or awkward sentence structure.  
*Example:* "I think we should part ways, but I still care about you and love you more than anything in the world."  
4 - The response is generally coherent, with only slight issues in phrasing or flow.  
*Example:* "I believe it’s time for us to end things, and I’m sorry if that hurts."  
5 - The response is fully clear and easy to follow, with smooth transitions between ideas.  
*Example:* "This decision was hard, but I feel that breaking up is the best thing for both of us."