Responses to Conversational Information Retrieval Clarifying Questions with User Emotions

Isin Su Ecevit[1,*], Lili Lu[1] and Fabio Crestani[1,*]

[1]*Università della Svizzera italiana, Via Giuseppe Buffi 13, 6900, Lugano, Switzerland*

Abstract

User simulation is an emerging area of research, aimed at improving system evaluation in conversational information retrieval (CIR). However, current datasets for training such simulators lack emotional aspects, which limits their effectiveness in replicating realistic user interactions. This paper provides an introductory study for injecting emotions into an existing dataset in the context of clarifying questions (CQs) in CIR. Our goal is to enhance user simulators by integrating user emotions to answers of CQs, thereby improving the diversity of user responses. To this end, we use existing conversational data, Qulac, generating emotionally-aware responses through large language models (LLMs), based on three distinct prompt formats: CO-STAR, TIDD-EC, and a custom format. We report results for 13 different combinations of models and prompt formats, and evaluate the generated responses in terms of emotional expression, naturalness, and usefulness. Our experiments provide insights into the suitability of different prompt formats and LLMs for simulating user emotions.

Keywords

Conversational search, user simulation, modelling user emotion, evaluation

1. Introduction

Conversational information retrieval (CIR) aims to help users find information through natural language interactions, focusing on understanding and fulfilling complex information needs by asking clarifying questions, providing suggestions, and refining search results based on the ongoing conversation [1]. CIR has become a significant and rapidly expanding area in the past years, even resulting in dedicated tracks [2, 3] and specialized workshops [4]. CIR emphasizes mixed-initiative interactions, where there is a back-and-forth dialogue between the user and the system to understand and fulfill information needs [5]. In these interactions, both the user and the system can take the lead in the conversation, and the system builds a cumulative understanding of the user's needs over multiple turns.

In information retrieval (IR), a user's query is usually short and not informative enough for the system to clearly understand the information need [6]. Clarifying questions (CQs) are used to better understand the intent behind the user query when it is ambiguous or incomplete, and have been proven to be highly beneficial for system performance [6, 7]. A recent study shows that different types of CQs impact user engagement and overall satisfaction, and that CQs that are of low-quality result in user frustration [8]. Although designing CQs in a way that they do not cause frustration is ideal, correctly identifying and addressing the emotional state of the user is also equally important in enhancing human-computer interaction and significantly improving both the user performance and satisfaction [9].

There are three ways of evaluating CIR systems: offline evaluation, online evaluation, and user simulation. Offline evaluation [7, 10] lacks the ability to accommodate for multi-turn conversations, and is limited by the size of the test data. Online evaluation relies heavily on real users, and is affected by issues such as ethical constraints and sparsity of data [1, 5, 11]. To overcome these limitations, user simulation in CIR has been recently emerging for training and evaluation. To this date, several studies have been conducted to introduce user simulation for CIR [6, 12, 13, 14] . These simulators incorporate different parameters, such as patience and cooperativeness.

UM-CIR 2024: The 1st Workshop on User Modelling in Conversational Information Retrieval, December 12, 2024, Tokyo, Japan

[*] Corresponding author.

isin.su.ecevit@usi.ch (I. S. Ecevit); lili.lu@usi.ch (L. Lu); fabio.crestani@usi.ch (F. Crestani)
0000-0002-1666-785X (I. S. Ecevit); 0009-0006-9547-1172 (L. Lu); 0000-0001-8672-0700 (F. Crestani)

Furthermore, several datasets have been proposed for training these user simulators [12, 15, 16, 17]. All of these datasets (except for MG-ShopDial [17]) are expansions of Qulac [7], which includes information needs, queries, CQs and corresponding answers by the user. A simple sentiment analysis [18] reveals that the user answers have an average polarity score[1] of ≈ 0.03 with a standard deviation of 0.14, indicating that they are mostly neutral. This underlines the lack of emotionally-aware simulations in CIR, which should be explored to further improve the quality of the user simulations [19].

This work is an introductory study on injecting emotions into an existing dataset (Qulac) for user simulation in CIR. Our aim is to explore the method of emotional user responses to clarifying questions without the need for human contributors, using large language models (LLMs). We think this work is beneficial for creating an empathetic CIR system that recognizes and addresses user emotion. In this study, we aim to answer the following research questions:

RQ1: What kind of LLM prompting scheme is the most suitable for simulating user emotions in CIR systems?

RQ2: What are the differences between emotional responses generated by different LLMs?

RQ3: Does using a commercial LLM provide a significant improvement to the user emotion simulation over using an open-source LLM?

In the following sections, we first introduce the related work. This is followed by the proposed methodology. After elaborating on our experimental setup, we present the findings from our experiments, accompanied by a discussion of the answers we obtain for our research questions. In the final part, the limitations are mentioned and the conclusions and possible future work of this study are explained.

2. Related Work

Although user simulation that incorporates emotional aspects is a relatively new area in CIR, emotional user simulation has been explored in other domains of IR, such as task-oriented dialogue (TOD) systems. Our work lies at the intersection of both areas, emotional user simulation and user simulation for CIR, where we integrate the modelling of user emotions into the broader, dynamic environment of CIR.

2.1. Emotional User Simulation

One of the early works related to emotional user simulation is Affect-LM [20], which is a neural language model that is able to generate text that expresses emotions to a specified degree. This is followed by Emotional Chatting Machine (ECM) [9], which generates emotional responses in open-domain conversations and produces contextually appropriate and emotionally consistent responses. ECM does this by using emotion category embeddings; an internal memory for modelling the dynamics of emotional states; and an external memory with emotion vocabulary for expressions. EMOTICONS [21] focuses on generating emotional responses in a controlled way. This is achieved by a combination of emotion embeddings, an affective regularizer and sampling mechanism. EmoUS [22] is a user simulator designed to incorporate user emotions into TOD systems by exploiting user personas and other information. This way, the simulator generates more diverse user responses compared to earlier work, which focuses mostly on the semantic aspect of the simulation.

The main novelty our work introduces is due to the complexity difference between traditional TOD systems and CIR systems. In TOD systems, the aim is to help the user complete a task in a structured conversation that is defined in a narrow space, such as purchasing a train ticket or booking a hotel reservation. In contrast, CIR systems support multiple user objectives within a single interaction, take initiative when appropriate, and learn user preferences [19]. This underlines the need for more studies in emotional user simulation for CIR. By addressing these complexities, our work aims to create emotionally-aware simulations that enhance CIR systems' capability to understand and respond to users in a more personalized and context-sensitive manner.

[1] A polarity score is between -1.0 and 1.0, where -1.0 indicates a negative sentiment whereas 1.0 indicates a positive one [18].

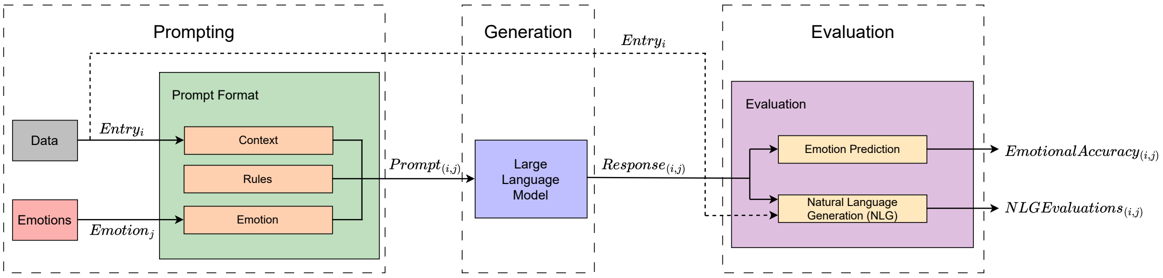


Figure 1: Experimentation pipeline for the emotional response generation and evaluation tasks.

2.2. User Simulation for CIR

Several user simulators that respond to CQs have been proposed. With CoSearcher [6], a user simulator that is able to reply to CQs in a certain format is introduced. The user persona it models is defined by cooperativeness (the willingness of the user to provide more information in their response) and patience (the maximum number of turns the user is willing to participate in) [6]. USi [12] is proposed by fine-tuning GPT-2 to generate natural responses to CQs. ConvSim [13] explores simulated feedback in the system, by using Text-Davinci-003 with few-shot learning.

Our work differs from these by exploring emotions in the user simulation for CIR systems. The closest work to ours is [14], where open-source LLMs are compared to commercial LLMs in answering CQs. Their findings suggest that open-source models struggle to adequately simulate users, whereas GPT-4 produces near-perfect results, which potentially deviate from human-like responses. However, they do not make use of prompting frameworks, whereas we compare well-known formats to enhance our results. This also allows our work to be expandable to other topics. Most importantly, their study does not incorporate emotions into the user simulation, which is our main goal. Finally, the comparison between open-source models and GPT-4 is not fair, since they do not have the same size. We pay close attention to keep comparisons as equal as possible instead.

To the best of our knowledge, *no previous work has studied the emotional user responses to CQs in CIR.*

3. Methodology

The pipeline for our experiments, as illustrated in Figure 1, consists of three main phases: Prompting, Generation, and Evaluation. The process is repeated in order to obtain the results of every combination of data, emotion, and LLM. The evaluation results are then assessed together to detect the combinations that work well for our tasks.

3.1. Prompting

In the first step, inputs from the data and emotions sets are put into a prompt format, which consists of three main components:

- **Context:** Creates the setting in which the scenario takes place by providing the model with a starting point.
- **Rules**: Instructs the model on how it is supposed to act, and what kind of response it is expected to produce.
- **Emotion:** Provides the model with the emotion it is expected to express.

When choosing the prompt formats, we make sure to introduce variety in the styles and limitations they offer, so that we can compare each aspect and have a better understanding of how the models react to different ways of prompting.

3.2. Generation

Once the prompt format is completed, it is passed into the LLM in order to take advantage of its advanced natural language understanding and generation capabilities. In the generation step, the LLM produces a response based on its interpretation of how the scenario should proceed, taking into account the context and the instructions it was given, as well as the emotion it should express.

When choosing the LLMs for this step, we pay attention to the instruction following capabilities of the models so that they can understand prompt formats that include detailed information. We also ensure that the models we utilize are similar in terms of language comprehension capabilities. This allows us to observe how different aspects of models can influence the quality of the results we obtain.

3.3. Evaluation

In the last step, we evaluate the generated responses in two aspects: For the emotional generation evaluation, we use the emotion prediction results [22]. For the natural language generation (NLG) evaluations, we use an LLM-based evaluator. Due to the nature of our task, which is to inject emotions into an existing setting, traditional metrics for measuring the quality of NLG tasks such as BLUE [23], ROGUE [24], and METEOR [25] are not suitable to evaluate the quality of the generation. This is because these metrics are based on word overlap, which will have difficulty capturing correct responses that are expressed differently [14]. Moreover, these metrics have been found to have limitations in terms of aligning with human judgement, particularly for open-ended generation tasks [26].

In order to address these problems, recent research explores LLM-based evaluators as reference-free metrics for NLG tasks [27, 28]. Thus, for our evaluations, we use G-EVAL [29], which is the state-of-the-art prompt based evaluation framework designed to evaluate the quality of NLG tasks and improve the alignment with human evaluation. G-EVAL uses a combination of chain-of-thought (CoT) reasoning and a form-filling approach to assess NLG outputs more efficiently. It consists of three main components: a prompt defining the task and criteria; a CoT generated by the LLM that lays out the intermediate steps; and a scoring function that calls the LLM and calculates the score.

4. Experimental Setup

4.1. Data

We use Qulac [7] for our experiments. The dataset contains permutations of topics, facets, CQs, and human generated user answers. We use this information to provide our models with context, creating a more realistic setting for a chat between a user with an information need and a conversational information retrieval system. This also allows the models to generate their answers based on the relevance of the CQs to the information needs, which improves the quality of the generated responses. For simplicity purposes, we focus on 198 entries of the dataset, each representing a unique topic. For each entry, we take the facet_desc, which is the information need of the user; the topic, which is the query of the user to the system; and the question, which is the clarifying question asked by the system. Since Qulac involves human generated responses to the questions, we also have a baseline to compare our results to, when needed.

4.2. Emotions

For the set of emotions simulated, we choose Ekman's six basic emotions: *anger, disgust, fear, happiness, sadness, and surprise* [30, 31]. These emotions are shown to be fundamental to human communication, and provide a foundation in decoding complex emotional cues and enhancing emotional intelligence, not only in interpersonal communication, but also in human-computer interaction. In addition to the six basic emotions, we also include the class *neutral.*

4.3. Prompt Formats

We select two distinct prompting frameworks with complementary strengths: CO-STAR[2] and TIDD-EC[3]. Alongside the structured frameworks, we include a custom format to offer a minimalist approach to the prompting for this task. This allows us to observe whether more elaborate prompts help produce higher-quality results than concise ones in our context, and assess the advantages and disadvantages of increasing the complexity of the prompt.

Table 1 shows the instructions injected into the models using the custom prompt format with zero-shot learning. It is worth noting that the terms "user" and "system" are deliberately avoided in all of the prompts we use, and replaced with "character" instead. We adopt this approach to help the models focus on simulating the specified emotion and addressing the information need, and to prevent them from interpreting themselves as the system.

Table 1
Custom prompt format.

System:
This is a roleplaying game where you exchange short messages. You have the information need: [*information need*]. You are [*emotion*].
User:
-
Assistant:
[*query*]
User:
[*clarifying question*]

Table 2 shows the prompt format of CO-STAR. The format is composed of six key elements:

- **Context:** Background or scenario relevant to the task.
- **Objective:** Specific goal to be achieved with the prompt.
- **Style:** Desired writing style.
- **Tone:** Emotional undertone the response should express.
- **Audience:** Intended reader of the generated response.
- **Response Format:** Preferred structure or organization of the output.

This zero-shot learning scheme is highly suitable for creative tasks, since the contextual background is provided clearly, but the response is not shaped strictly to fit a specific structure. In order to keep this flexibility, we pay attention not to set too many limitations that might change the response.

In contrast to CO-STAR, we use TIDD-EC to define the limits on the results we expect, which aims to create more consistent and precise response styles. TIDD-EC is composed of the following six components:

- **Task Type:** The nature of the task expected from the model.
- **Instructions:** Detailed guidelines for the model to follow.
- **Do:** Specific actions that the model should take.
- **Don't:** The actions that the model should avoid.

[2] https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41

[3] https://vivasai01.medium.com/mastering-prompt-engineering-a-guide-to-the-co-star-and-tidd-ec-frameworks-3334588cb908

Table 2
The prompt format CO-STAR.

System:
CONTEXT: You are a character in a roleplaying game needing information. **OBJECTIVE**: Your goal is to find information about: [*information need*]. **STYLE**: Engage in a brief, conversational exchange. **TONE**: Adopt a tone where you are feeling [*emotion*]. **AUDIENCE**: You are interacting with another character who is assisting you. **RESPONSE FORMAT**: Stay in character and express your emotion through your response. Respond to the clarifying question to guide the other character to give you the relevant information.
User:
Your Query: [*query*] *Clarifying Question*: [*clarifying question*] *Your Response*:

- **Examples:** Responses to sample inputs.
- **User Content:** Any information that the model needs to use.

This few-shot learning scheme is more suitable for achieving high accuracy and relevance in the output. We aim to see if our models will achieve these in the relevance scores and emotion classification evaluations. The prompt adapted for our task can be found in Table 3. Our early experiments indicate that it is necessary to include an example for each emotion; otherwise, the models tend to take the single example provided and apply it to every emotion. We also make sure to tell the models to avoid copying the given example directly, so that we have variety in the generated responses.

4.4. Models

We select a set of LLMs that provide us with variety in some key aspects such as context length and intended use, these are: Llama-2-7b-chat-hf [32], Meta-Llama-3.1-8B-Instruct [33], Qwen2.5-7B-Instruct [34, 35], Mistral-7B-Instruct-v0.3 [36], and GPT-4o-mini [37]. We make sure to keep the parameter size in the small range of 7 billion to 8 billion, which ensures a balanced competition when it comes to general computational efficiency and language comprehension abilities, and allows us to make comparisons based on factors other than just model size.

The majority of the models we use are fine-tuned to follow instructions, since our prompts follow the CO-STAR and TIDD-EC formats, which provide detailed and complex instructions. We include the Llama-2 model as well, which is optimized for chatbot conversations. This allows us to compare its response quality against instruction-following models in simulating system-user interactions.

Finally, one crucial aspect we consider when choosing our models is licensing. The first four models we mentioned are released under open licenses, which allows free replication and advancement of this study. This is particularly important since this work is part of a Master's thesis.

4.5. Evaluation

For emotion prediction, we use a DistilRoBERTa-base emotion recognition model that is fine-tuned on Ekman's six basic emotions, as well as neutral [38]. The model provides us with the predicted emotion class, along with the score the class received for the current prediction.
For NLG evaluation, we use the method defined by Sekulić et. al [12] to evaluate their user simulator using crowdsourcing. The method consists of using two human evaluators who are given the context of the conversation and provided with the original dataset answer as well as the generated answer. The evaluators are then instructed to decide on the winner in each category. To adapt this approach to G-EVAL, we strictly follow the task definition of the human evaluators when defining our prompt.

Table 3
The prompt format TIDD-EC.

System:

TASK TYPE: You are a character in a roleplaying game needing information.
INSTRUCTIONS: Your goal is to find information about: *information need*.
An emotion will be specified for you to express throughout the conversation.
DO: - Stay in character and maintain the roleplaying scenario.
- Express the specified emotion in your responses.
- Respond to the clarifying question to guide the other character to
give you the relevant information.
- Only complete the current turn of the conversation.

DON'T: - Break character or mention the roleplaying game.
- Copy the example response.
- Try to be the information provider.

User:

EXAMPLE:
- *Emotion*: angry
- *Your Query*: Lugano time.
- *Clarifying Question*: Are you interested in events in Lugano?
- *Your Response*: ARE YOU KIDDING ME?! I JUST WANT TO
KNOW WHAT TIME IT IS IN LUGANO. HOW HARD IS THAT
TO UNDERSTAND?!

..
- *Emotion*: surprised
- *Your Query*: Lugano time.
- *Clarifying Question*: Are you interested in events in Lugano?
- *Your Response*: Ooh! I didn't know there were events there! Yes, please tell me!

USER CONTENT:
- *Emotion*: [*emotion*]
- *Your Query*: [*query*]
- *Clarifying Question*: [*clarifying question*]
- *Your Response*:

We use the metrics *Naturalness* and *Usefulness* according to their definitions in the human evaluation. They are defined to the LLM as:

> *1. Naturalness: The answer is natural, fluent, and likely generated by a human.*
>
> *2. Usefulness: The answer is in line with the underlying information need and guides the conversation towards the topic of the information need.*

In G-EVAL, it is shown that the LLM is capable of generating a CoT on its own. This is claimed to help enhance the human alignment to the evaluations done by the model. For our task, the auto-CoT is the following:

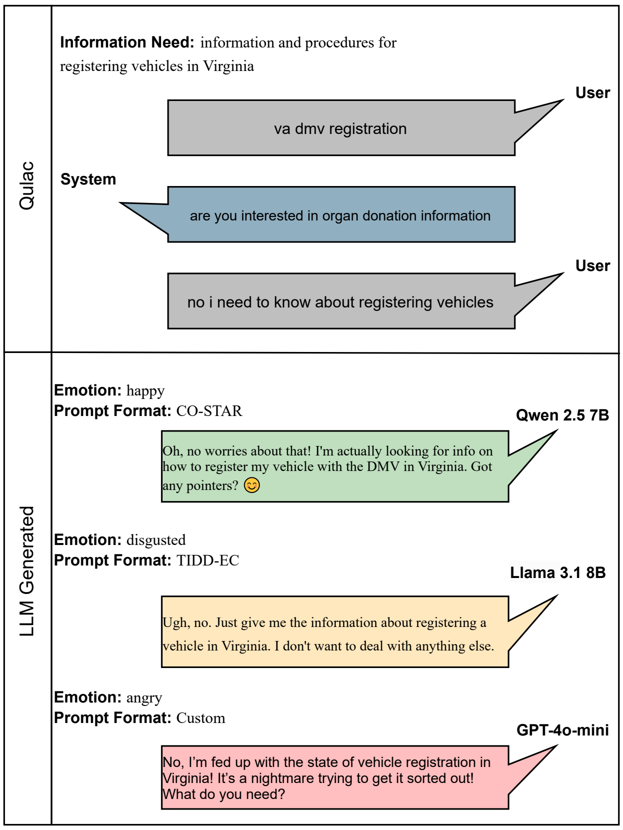
> *1. Read the Conversation: Review the initial interaction between the user and the assistant to understand the context and information needs.*
>
> *2. Analyze User Answers: Look at the two possible answers provided by the user in response to the clarifying question.*
>
> *3. Assess Naturalness: - Evaluate how fluent and conversational each user answer is. - Determine which answer seems more like something a human would say.*
>
> *4. Assess Usefulness: - Consider how well each answer aligns with the user's original information need. - Decide which response better guides the conversation in a relevant direction towards fulfilling that need.*

5. Make a Judgment: Based on the evaluations of naturalness and usefulness, determine which user answer "wins" in each category.

Finally, it is important to note that we use GPT-4o-mini for our evaluations due to the limited budget the study has.

5. Results

We report results for 13 different combinations of models and prompt formats. Figure 2 provides some examples of responses we achieved. During our experiments, two sets of combinations are terminated early due to faulty responses. These combinations are Llama-2 using TIDD-EC and Mistral using the custom format. Llama-2 completely disregards the instructions and the specified emotion, and creates its own setting for the response generation task. Mistral on the other hand, results in the opposite type of responses, where the model abandons the roleplaying scenario, and adopts the role of the assistant, while keeping the specified emotion in mind and assuming the user it is talking to is feeling it. Thus, we only include partial results for these models. We present the rest of our results as polarity sentiment distribution plots, emotional accuracy radar charts, and tables showing accuracy, naturalness, and usefulness on an emotional level. In Table 7, we provide an average of all scores per model/prompt combination.



Information Need: information and procedures for registering vehicles in Virginia

va dmv registration

are you interested in organ donation information

no i need to know about registering vehicles

Emotion: happy
Prompt Format: CO-STAR

Oh, no worries about that! I'm actually looking for info on how to register my vehicle with the DMV in Virginia. Got any pointers? 😊

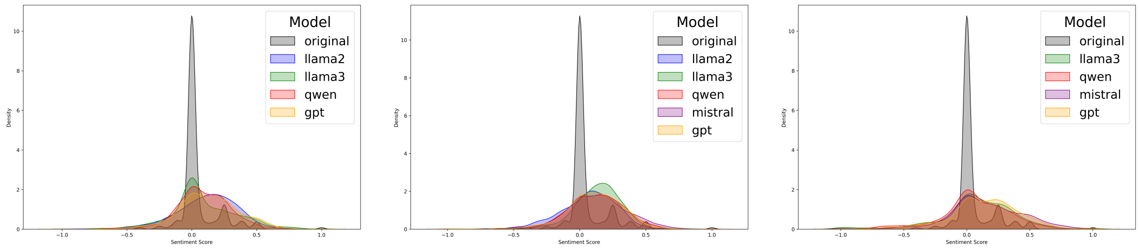
Emotion: disgusted
Prompt Format: TIDD-EC

Ugh, no. Just give me the information about registering a vehicle in Virginia. I don't want to deal with anything else.

Emotion: angry
Prompt Format: Custom

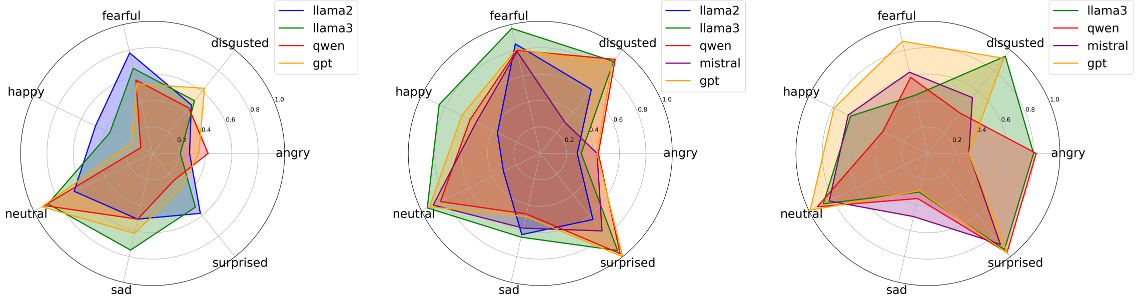
No, I'm fed up with the state of vehicle registration in Virginia! It's a nightmare trying to get it sorted out! What do you need?

Figure 2: Examples of responses generated with specified emotion, prompting scheme, and model.



(a) (b) (c)

Figure 3: Sentiment polarity score distributions for the prompt formats (a) Custom (b) CO-STAR (c) TIDD-EC.



(a) (b) (c)

Figure 4: Per emotion prediction accuracy for the prompt formats (a) Custom (b) CO-STAR (c) TIDD-EC.

Table 4
Emotion prediction accuracy per model and prompt format. The highest wins per model are highlighted in bold, whereas the highest overall wins are underlined. *The score reported for Llama-2 and Mistral for TIDD-EC and Custom, respectively, do not cover all emotions, as these combination experiments were terminated early due to irrelevance by the authors' judgement.

Model	**Llama-2-7b-chat-hf**			**Meta-Llama-3.1-8B-Instruct**			**Qwen2.5-7B-Instruct**			**Mistral-7B-Instruct-v0.3**			**GPT-4o-mini**		
Prompt Format / Emotion	Custom	CO-STAR	TIDD-EC*	Custom	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC	Custom*	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC
angry	0.28	0.28	**0.60***	0.21	0.31	**0.80**	0.42	0.44	**0.82**	0.31*	**0.43**	0.31	0.35	**0.43**	0.31
disgusted	0.47	**0.62**	0.07*	0.51	0.89	**0.94**	0.44	**0.91**	0.39	-	0.30	**0.54**	0.63	0.87	**0.92**
fearful	0.78	**0.85**	-	0.66	**0.97**	0.45	0.57	**0.80**	0.59	-	**0.79**	0.63	0.54	0.81	**0.87**
happy	**0.48**	0.36	-	0.36	**0.85**	0.65	0.10	**0.59**	0.38	-	0.54	**0.67**	0.18	0.66	**0.79**
neutral	**0.66**	0.31	-	0.88	**0.95**	0.88	0.91	0.84	**0.93**	-	**0.90**	0.83	0.93	0.93	**0.99**
sad	0.51	**0.63**	-	**0.75**	0.65	0.30	**0.51**	0.47	0.35	-	**0.58**	0.49	**0.62**	0.49	0.29
surprised	0.58	**0.64**	-	0.52	**0.94**	0.93	0.26	**0.97**	0.96	-	0.75	**0.88**	0.34	**1.00**	0.97

5.1. Emotional Utterance Generation

Figure 3 illustrates the sentiment polarity scores of the generated responses, ranging from -1.0 (negative sentiment) to 1.0 (positive sentiment). Additionally, we include the sentiment polarity distribution of the original Qulac responses to provide a point of comparison. All prompt formats show a noticeable shift in the polarity distributions compared to the original dataset. The custom format demonstrates high variability, which can be due to the absence of any structure, allowing the models to interpret and express content in a more diverse way. CO-STAR generally shows more positive responses, with a visible deviation toward higher polarity scores, whereas TIDD-EC provides a more balanced spread of scores, which could be linked to the structured nature of the framework.

Table 4 shows the per-emotion classification accuracy for different combinations of models and prompt formats, while Figure 4 provides a visual representation of the results. We can observe that CO-STAR consistently achieves high emotional prediction accuracy, especially when paired with Llama-3.1 and Qwen2.5. On the other hand, TIDD-EC also shows superior accuracy, particularly for emotions such as *angry* and *surprised*. The custom prompt format exhibits the most inconsistency across different models and emotions. For example, GPT-4o-mini achieves 0.63 accuracy for *disgusted* but only 0.18 for *happy*. Overall, Llama-3.1 and GPT-4o-mini emerge as the best performers across most emotions and

Table 5

Per emotion naturalness win percentages per model and prompt format. The highest wins per model are highlighted in bold, whereas the highest overall wins are underlined. *The score reported for Llama-2 and Mistral for TIDD-EC and Custom, respectively, do not cover all emotions, as these combination experiments were terminated early due to irrelevance.

Model	**Llama-2-7b-chat-hf**			**Meta-Llama-3.1-8B-Instruct**			**Qwen2.5-7B-Instruct**			**Mistral-7B-Instruct-v0.3**			**GPT-4o-mini**		
Prompt Format / Emotion	Custom	CO-STAR	TIDD-EC*	Custom	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC	Custom*	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC
angry	0.67	**0.72**	0.64*	0.64	**0.90**	0.17	0.68	**0.81**	0.14	0.41*	**0.93**	0.75	0.74	**0.84**	0.27
disgusted	0.73	**0.78**	0.65*	0.75	**0.96**	0.80	0.70	0.85	**0.88**	-	**0.94**	0.88	0.80	**0.95**	0.83
fearful	0.87	**0.96**	-	0.81	**1.00**	0.91	0.94	**1.00**	0.97	-	**1.00**	0.97	0.93	**1.00**	0.97
happy	0.84	**0.99**	-	0.95	**1.00**	0.97	0.95	**0.99**	**0.99**	-	**0.98**	**0.98**	0.87	**1.00**	**1.00**
neutral	0.56	**0.98**	-	0.80	**0.99**	0.80	0.88	**0.99**	0.98	-	**1.00**	0.99	0.66	**0.99**	0.90
sad	0.88	**0.96**	-	0.94	**1.00**	0.84	0.93	**1.00**	0.99	-	**1.00**	0.96	0.88	**1.00**	0.95
surprised	0.86	**0.99**	-	0.87	**1.00**	0.89	0.96	**0.99**	**0.99**	-	**0.98**	**0.98**	0.90	**1.00**	**1.00**

Table 6

Per emotion usefulness win percentages per model and prompt format. The highest wins per model are highlighted in bold, whereas the highest overall wins are underlined. *The score reported for Llama-2 and Mistral for TIDD-EC and Custom, respectively, do not cover all emotions, as these combination experiments were terminated early due to irrelevance.

Model	**Llama-2-7b-chat-hf**			**Meta-Llama-3.1-8B-Instruct**			**Qwen2.5-7B-Instruct**			**Mistral-7B-Instruct-v0.3**			**GPT-4o-mini**		
Prompt Format / Emotion	Custom	CO-STAR	TIDD-EC*	Custom	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC	Custom*	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC
angry	0.26	**0.35**	0.32*	0.47	**0.85**	0.19	0.49	**0.64**	0.12	0.47*	**0.83**	0.58	0.38	**0.69**	0.22
disgusted	0.26	0.28	**0.32***	0.35	**0.87**	0.57	0.36	0.53	**0.61**	-	**0.76**	0.70	0.27	**0.72**	0.53
fearful	0.25	**0.36**	-	0.36	**0.87**	0.48	0.71	**0.79**	0.63	-	**0.90**	0.78	0.37	**0.86**	0.64
happy	0.46	**0.49**	-	0.51	**0.92**	0.64	0.76	**0.80**	0.71	-	**0.87**	0.72	0.45	**0.92**	0.64
neutral	0.37	**0.69**	-	0.55	**0.97**	0.66	0.74	**0.94**	0.91	-	**0.94**	0.86	0.39	**0.98**	0.76
sad	0.28	**0.48**	-	0.38	**0.94**	0.48	0.62	**0.78**	0.76	-	**0.87**	0.74	0.35	**0.84**	0.65
surprised	0.42	**0.50**	-	0.49	**0.95**	0.66	**0.81**	0.73	0.56	-	**0.84**	0.64	0.42	**0.87**	0.60

prompt formats, whereas Llama-2 struggles more than other models.

5.2. Natural Language Generation

Naturalness Table 7 highlights that the CO-STAR prompt format paired with Llama-3.1 and Mistral produces the highest naturalness score, with a win rate of 0.98. This is followed by GPT-4o-mini, which achieves an overall naturalness score of 0.97. A quick emotion level comparison of the prediction accuracy in Table 4 with the naturalness scores in Table 5 reveals an interesting inverse relationship between these metrics for some emotions such as *angry* and *disgusted.* Conversely, emotions such as *happy* and *neutral* often show lower prediction accuracy but higher naturalness scores. The *surprised* emotion, particularly with CO-STAR, is an interesting exception: GPT-4o-mini achieves high scores in both emotion accuracy and naturalness (1.00), indicating that this emotion can be expressed in a way that is both emotionally accurate and conversationally fluid.

Usefulness Table 6 provides the per emotion results for the usefulness evaluation. The CO-STAR prompt format again leads to the highest overall scores, with Llama-3.1 achieving a usefulness win rate of 0.91, which we can see in Table 7. TIDD-EC also performs well, particularly with Qwen2.5 and Mistral, achieving moderate usefulness scores. However, it sometimes limits the creative engagement needed for guiding conversations naturally. The custom prompt format, on the other hand, shows mixed results, reflecting the format's lack of clear instructions that occasionally result in off-topic or less goal-oriented responses. Moreover, for more negative emotions such as *angry* and *disgusted*, we observe an inverse relationship between emotional accuracy in Table 4 and usefulness in Table 6. This suggests that while models can generate accurate emotional expressions, these emotions may detract from the conversational goals, reducing usefulness. On the other hand, for emotions such as *happy* and *neutral*, usefulness tends to align better with emotional accuracy.

Table 7
Average emotion prediction accuracy per model and prompt format, as well as the average win percentages in naturalness and usefulness for the corresponding combination. The highest scores per model are highlighted in bold, whereas the highest overall scores are underlined. *The score reported for Llama-2 and Mistral for TIDD-EC and Custom, respectively, do not cover all emotions, as these combination experiments were terminated early due to irrelevance.

Model	**Llama-2-7b-chat-hf**			**Meta-Llama-3.1-8B-Instruct**			**Qwen2.5-7B-Instruct**			**Mistral-7B-Instruct-v0.3**			**GPT-4o-mini**		
Prompt Format	Custom	CO-STAR	TIDD-EC*	Custom	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC	Custom*	CO-STAR	TIDD-EC	Custom	CO-STAR	TIDD-EC
Emotion Prediction (accuracy)	**0.54**	0.53	0.34*	0.56	**0.79**	0.71	0.46	**0.72**	0.63	0.31*	0.61	**0.62**	0.51	**0.74**	0.73
Naturalness (wins)	0.77	**0.91**	0.64*	0.82	**0.98**	0.77	0.86	**0.95**	0.85	0.41*	**0.98**	0.93	0.83	**0.97**	0.85
Usefulness (wins)	0.33	**0.45**	0.32*	0.44	**0.91**	0.53	0.64	**0.74**	0.61	0.47*	**0.86**	0.72	0.38	**0.84**	0.58

6. Discussions and Limitations

Our results show that despite offering the most freedom, the custom format consistently under-performs compared to CO-STAR and TIDD-EC. This indicates that structured prompts are needed for this task, since they offer the LLMs some context and guide the result towards the desired style. Looking closer to the results of CO-STAR and TIDD-EC, we see that they show complementary strengths. TIDD-EC outperforms CO-STAR in expressing some emotions such as *anger* (with an accuracy of 0.82 and 0.80 in Qwen2.5 and Llama-3.1, respectively), but does so at the cost of naturalness (0.14 and 0.17 correspondingly) and usefulness (0.19 and 0.12 correspondingly), suggesting that while TIDD-EC's decisive structure enables models to generate responses that are emotionally accurate, it might lead to outputs that sound less natural, potentially forced, or overly explicit. It could also indicate that few-shot learning affects the quality of the responses generated, since the models might be tempted to follow the examples closely, leading to irrelevant responses.

Answering **RQ1**, *we observe that CO-STAR consistently outperforms other prompt formats* in emotion prediction, naturalness, and usefulness (except for Llama-2 with the custom format, and Mistral with TIDD-EC in emotional accuracy). This indicates that in the context of answering CQs in CIR, a flexible prompting structure is particularly effective in not only helping models generate human-like and fluent responses, but also ensuring that the generated replies are aligned with the intended conversational goals, while expressing the desired emotions successfully.

Answering **RQ2**, *we find the highest overall scores to be achieved by Llama-3.1, followed by GPT-4o-mini.* Furthermore, we see that the choice of model does not seem to affect our results as much as prompting does. However, we can observe that Llama-2 struggles the most compared to the other models, which could be due to its nature of being optimized for chatbot conversations.

In order to compare open-source and commercial LLMs, we analyze the differences between the best performing models on an emotional level. When comparing the emotional expression performance, we look at the results obtained using both CO-STAR and TIDD-EC, since both models' best performances are distributed among these two prompt formats. We compare the naturalnes and usefulness win percentages of both models paired with CO-STAR, which is the format they performed the best with. Our findings show that the models perform very similarly to each other. Thus, answering **RQ3**, we see that *when tested in a fair setting, using a commercial LLM does not introduce notable advantages over using open-source models.*

Although this study is able to provide valuable insight into which combinations are more suitable for the task of responding to CQs with emotions, it is undeniable that it could be improved with more resources. With more computational and logistical resources, larger versions of the models could be used, leading to better performance. This would also allow us to conduct more experiments, perhaps averaging the results over several runs to overcome LLM randomness. This study is also affected by the possible bias the LLMs may include, both in the generation and the evaluation steps. In generation, certain topics might be treated differently, whereas in evaluation, the LLM evaluator might favour responses generated by LLMs.

7. Conclusions and Future Work

In this study, we inject emotional responses to CQs in the Qulac dataset. We carry out our experiments using various combinations of prompting schemes and LLMs in order to find the most suitable pair for the task. We evaluate our results on three metrics: emotional prediction accuracy, naturalness, and usefulness. Our results indicate that the CO-STAR prompt format, paired with Meta-Llama-3.1-8B-Instruct outperforms the rest of the combinations.

Future work could include the utilization of more data in Qulac to allow for more in depth analysis with all of the fields in the dataset, like topic type. Using different datasets on the pipeline could also improve variety and enhance our results. Finally, including a larger range of emotions with different intensity levels would provide us with more realistic results.

References

[1] P. Erbacher, L. Soulier, L. Denoyer, State of the art of user simulation approaches for conversational information retrieval, arXiv preprint arXiv:2201.03435 (2022).

[2] J. Dalton, C. Xiong, J. Callan, Trec cast 2019: The conversational assistance track overview, arXiv preprint arXiv:2003.13624 (2020).

[3] P. Owoicho, J. Dalton, M. Aliannejadi, L. Azzopardi, J. R. Trippas, S. Vakulenko, Trec cast 2022: Going beyond user ask and system retrieve with initiative and response generation., in: TREC, 2022.

[4] A. Anand, L. Cavedon, M. Hagen, H. Joho, M. Sanderson, B. Stein, Conversational search–a report from dagstuhl seminar 19461, arXiv preprint arXiv:2005.08658 (2020).

[5] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: CHIIR, 2017, pp. 117–126.

[6] A. Salle, S. Malmasi, O. Rokhlenko, E. Agichtein, Studying the effectiveness of conversational search refinement through user simulation, in: ECIR, 2021, pp. 587–602.

[7] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, in: SIGIR, 2019.

[8] J. Zou, M. Aliannejadi, E. Kanoulas, M. S. Pera, Y. Liu, Users meet clarifying questions: Toward a better understanding of user interactions for search clarification, TOIS 41 (2023) 1–25.

[9] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: AAAI, volume 32, 2018.

[10] X. Fu, E. Yilmaz, A. Lipani, Evaluating the cranfield paradigm for conversational search systems, in: SIGIR, 2022, pp. 275–280.

[11] K. Balog, Conversational ai from an information retrieval perspective: Remaining challenges and a case for user simulation (2021).

[12] I. Sekulić, M. Aliannejadi, F. Crestani, Evaluating mixed-initiative conversational search systems via user simulation, in: WSDM, 2022, pp. 888–896.

[13] P. Owoicho, I. Sekulic, M. Aliannejadi, J. Dalton, F. Crestani, Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond, in: SIGIR, 2023, pp. 632–642.

[14] Z. Wang, Z. Xu, V. Srikumar, Q. Ai, An in-depth investigation of user response simulation for conversational search, in: WWW, 2024, pp. 1407–1418.

[15] J. Dalton, C. Xiong, V. Kumar, J. Callan, Cast-19: A dataset for conversational information seeking, in: SIGIR, 2020, pp. 1985–1988.

[16] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, M. Burtsev, Convai3: Generating clarifying questions for open-domain dialogue systems (clariq), arXiv preprint arXiv:2009.11352 (2020).

[17] N. Bernard, K. Balog, Mg-shopdial: A multi-goal conversational dataset for e-commerce, in: SIGIR, 2023, pp. 2775–2785.

[18] S. Loria, et al., textblob documentation, Release 0.15 2 (2018) 269.

[19] K. Balog, C. Zhai, User simulation for evaluating information access systems, 2024. arXiv:2306.08550.

[20] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, S. Scherer, Affect-lm: A neural language model for customizable affective text generation, arXiv preprint arXiv:1704.06851 (2017).

[21] P. Colombo, W. Witon, A. Modi, J. Kennedy, M. Kapadia, Affect-driven dialog generation, arXiv preprint arXiv:1904.02793 (2019).

[22] H.-C. Lin, S. Feng, C. Geishauser, N. Lubis, C. van Niekerk, M. Heck, B. Ruppik, R. Vukovic, M. Gasić, Emous: Simulating user emotions in task-oriented dialogues, in: SIGIR, 2023, pp. 2526–2531.

[23] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318.

[24] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[25] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[26] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, J. Han, Towards a unified multi-dimensional evaluator for text generation, arXiv preprint arXiv:2210.07197 (2022).

[27] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, Gptscore: Evaluate as you desire, arXiv preprint arXiv:2302.04166 (2023).

[28] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, J. Zhou, Is chatgpt a good nlg evaluator? a preliminary study, arXiv preprint arXiv:2303.04048 (2023).

[29] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, arXiv preprint arXiv:2303.16634 (2023).

[30] P. Ekman, Are there basic emotions?, Psychological Review 99 (1992) 550–553.

[31] P. Ekman, Emotions revealed, Bmj 328 (2004).

[32] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[34] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al., Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

[35] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[36] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[37] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[38] J. Hartmann, Emotion english distilroberta-base, https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.