Appendix for the paper: Socio-Emotional Response Generation: A Human Evaluation Protocol for LLM-Based Conversational Systems

I. Appendix A: Choice of the Next Label Prediction Model

In this first step, we aim to evaluate the performance of various models on the task of predicting a sequence of labels that models the social and emotional behaviours that are expected to be displayed in a generated response to a conversational context. In other words, we want to test the first step of our approach and determine the most suitable model to use as the planning module.

a) Data Preprocessing: We work with the Daily Dialog dataset. For each speaker turn, we consider 3 dialogue turns as the "context" and pair them with the label(s) of the following utterance to constitute a training sample. The model thus learns how to predict the labels of the next speaker turn. Our resulting train/validation/test splits are made up of 76052 / 7070 / 6740 samples.

Utterance	**Labels**
You surely know a lot about Chinese tea.	inform
Sure, I like drinking tea at teahouses.	inform, happiness
Oh, so do I.	inform
Why don't we go for one now?	directive
Great. We can chat while enjoying a cup there.	commissive, happiness

For such a conversation, we can extract a few training samples, for example:
Context: 'Sure, I like drinking tea at teahouses. — Oh, so do I. — Why don't we go for one now ?'
Labels: 'commissive, happiness'

b) Models: All the models we present were trained using a single GPU (NVIDIA RTX 8000, 48GB memory), with the hyper-parameters described in the Appendix **??**. We describe the models used in Experiment 1 below:

c) BERT - Multilabel Classification: BERT Base (110M parameters) and *BERT Large* (340M parameters) [1] are trained on a multi-label classification task. We set the confidence threshold at 0.7 for BERT Base and 0.5 for BERT Large.

d) BART - Sequence Generation: BART Base (140M parameters) and *BART Large* (406M parameters) [2] are fine-tuned on the task of generating the next labels sequence.

e) Beluga - Prompt-Based Generation: We use Beluga (13B parameters), a Llama2 model [3] fine-tuned on an Orca style dataset, to generate the sequence of the next labels using few-shots prompt-based generation. *Beluga* was prompted to generate the sequence of labels associated with the following speaker turn, given a dialogue utterance. The prompt used is:
Predict the sequence of labels associated with the utterance that follows the given dialogue.
We consider the following labels: 'inform', 'question', 'directive', 'commissive', 'neutral', 'anger', 'disgust', 'fear', 'happiness', 'sadness' and 'surprise'. The answer must be one or a sequence of multiple labels from this list.

Here are a few examples,
Dialogue: Good morning, sir. Is there a bank near here ?
Labels: 'inform'.
Dialogue: Is it far ?
Labels:'inform'
Dialogue: No, It's only about five minutes walk.
Labels: 'inform', 'happiness'.

What labels are associated with the utterance following this dialogue:
Dialogue: + [current utterance]

Random Selector - Baseline We add a random selector that will, for each utterance, select random labels out of the list of available labels. This model is meant to serve as a comparison with the other two models. We randomly select k labels out of the list, k chosen randomly between 1 and 2, following the length distribution observed in the dataset.

f) Metrics: As a sub-task of the response generation process, label sequence prediction is a one-to-many problem: many sequences can match a same context. However, efficiently evaluating the relevance of a sequence of labels to a context remains a challenging task due to the lack of suitable metrics. Thus, to evaluate this experiment, we must rely on comparing the pairs of sequences: the generated or predicted sequence, and the expected sequence.

To evaluate the results, we rely on metrics implemented in the *scikit-learn* library, such as the *Jaccard Score*, used to compare two sets of labels to evaluate the similarity between the predicted set and the expected set, or the multi-label implementation of *F1 score, Precision, Recall* that allows us to measure the performance of the models against the

expected sets of labels. For the sequence generation task (BART and Beluga), we also measure the *Normalised Levenshtein Similarity (NLS)* Levenshtein Distance (LD), a lexical similarity measure which identifies the distance between one pair of strings. It represents the smallest number of base edit operations, namely insertion, deletion or substitution, required to transform the source sequence S into the target sequence T. Levenshtein Similarity (LS) is computed as $LS = 1-LD$, and it is normalised as $NLS = (1 - LD)/max(len(T), len(S))$. Normalised Levenshtein Similarity is implemented in the *textdistance* library. Lastly, we look at the mean length of the generated/predicted sequences, *Mean* l_i, to contrast it with the dataset's average of 1.20 labels per utterance.

These metrics are efficient in comparing the gap between what is predicted and what labels were used in the real conversation, but it is important to keep in mind that when it comes to dialogue there is not one single good answer. There are many different ways to participate in a conversation, and there is no guarantee that a different agent would have used the same strategies.

As a classifier, BERT operates without inherent awareness of sequence order; it processes input as unordered lists rather than predictive sequences. While we did expect a lower performance, we believed it was interesting to compare a "safer" method, such as classification, that is forced to predict real labels, to generative methods that can be prone to hallucinating. However, when we look into the predictions outputted by the BERT models, we see that it only predicts the main class and does not manage to provide diverse outputs.

As for the generative approaches, the amount of data required to confidently fine-tune a generation model such as BART is quite demanding, especially on a task such as next utterance labels generation. Prompt-based approaches such as Beluga, through the use of few-shot prompting, offer a more data-efficient approach. We are interested to see if the capabilities of a large, Llama2-like model, can bridge the performance gap with a data-driven model such as BART. BART yields results more interesting and diverse than BERT's. While BART Base displays the best performance, the performance of BART Large does not parallel its larger scale, presenting comparatively inferior results. As prompt-based models have been rising with the success of ChatGPT, new possibilities have become accessible. However, when it comes to sensible or confidential data, it is hard to use such online services. We picked the Beluga model because it is an open-source alternative, fine-tuned from Llama 2, with good overall performances in English. However, we tried different prompts for Beluga but none were conclusive. The results are far below the previous methods, and even comparable to the low Random model baseline. One of the biggest issues with the prompt-based approach is that many results were not parsable, outputting 'None' as a label sequence even when explicitly told not to do so.

II. Appendix B: Details on the models used for Conditional Generation

Here are the main hyper-parameters used to train each model presented in this paper. Each model was trained using a single GPU (NVIDIA RTX 8000, 48GB memory).

A. Beluga: Prompts for Conditional Response Generation

Multiple prompts were tested to optimise the results and here are the final instructions used to train Beluga for the two experiments. Here, N is the number of sequences to be generated. In this paper, we used N = 10. Element refers to the dialogue history considered, we use a window of 3 utterances of context. We set the dialogue history in the format: SPEAKER A: utt1 SPEAKER B: utt2 SPEAKER A: utt3.

a) Beluga F&R: **For the generation of a single response, 'NO-CD' task, the prompt used is:** Generate the response following the given context.

For example:
A: Do you like some soup?
B: Yes, but I don't know what soup you have
A: We have beef soup and tomato soup
Response: Good. I prefer beef soup .

A: Can I take your order now, Madam?
B: Yes, what would you recommend?
A: I'm happy to recommend the fish, It tastes delicious, and it is today's special. Our chef is from the coast, and loves seafood. Today's special is actually his favorite dish. so I'm sure it is a
Response: It does sound wonderful, maybe I'll try it .

Generate the response following the following dialogue: + element

For the multiple responses generation, CD-pred and CD-GT tasks, the prompt used is: Generate + str(k) + responses following this dialogue: + element
Number the generated sequences from 1 to + str(k)
Generated sequences:
1:

b) Beluga PB: In this case, 'element' still stands for the 3-turn context, and 'labels' is the sequence of expected labels (*e.g. ''inform, happiness'*). The expected labels can either come from the dataset (task CD-GT) or from the prediction of a BART generative model (task CD-pred). The prompt used is:
Generate the response following the given context : + element
The tone of the response must be + labels
Response:

Model	Jaccard Score	Preci-sion	Recall	F1 Score	NLS	mean l_i
$BERT_b$	0.34	0.43	0.62	0.49	*NA*	0.58
$BERT_L$	0.38	1.00	0.38	0.55	*NA*	1.00
$BART_b$	0.38	0.56	0.53	0.54	0.54	1.15
$BART_L$	0.38	0.54	0.54	0.54	0.53	1.22
Beluga	0.020	0.04	0.05	0.04	0.099	2.72
Random	0.035	0.11	0.10	0.07	0.12	**1.20**

TABLE I
COMPARATIVE RESULTS OF THE EXPERIMENTS ON CONDITIONING THE GENERATION OF A MULTI-LABEL SEQUENCE OF SOCIAL AND EMOTIONAL BEHAVIOURS. $_B$ DENOTES A BASE MODEL, AND $_L$ INDICATES A LARGE MODEL.

Model	Trained Epochs	Learning rate	Batch-size
BERT			
Bert Base	10	3e-5	32
Bert Large	10	3e-5	32
Bert Current	20	3e-5	32
BART			
BART Base	10	3e-5	32
BART Large	10	3e-5	32

TABLE II
HYPER-PARAMETERS FOR TRAINING.

Model	Epochs	Learning rate	Batch-size
GPT-2			
GPT-2 Small	12	2e-5	32
GPT-2 Medium	5	2e-5	16
DialoGPT			
DialoGPT Small	8	2e-5	32
DialoGPT Medium	4	2e-5	16
BART		2e-5	
BART Base	10	2e-5	32
BART Large	6	2e-5	16

TABLE III
HYPER-PARAMETERS FOR TRAINING.

III. APPENDIX C: GENERATION EXAMPLE

Let's consider the following dialogue:
SPEAKER A: Good morning. What's the matter with you?
SPEAKER B: Good morning, Doctor. I have a terrible headache.
SPEAKER A: All right, Young man. Tell me how it got started.

The ground truth label is 'inform'. The labels generated by the BART model are ['inform', 'commissive']. The actual response to the dialogue is:

ACTUAL RESPONSE: Yesterday I had a runny nose. Now my nose is stuffed up. I have a sore throat. And I'm afraid I've got a temperature. I feel terrible.
The following table IV presents the answers generated by all the considered models and approaches.

Model	Response
GPT-2	
GPT-2	It started on me at work and has continued on me at home. I am having a bad headache in the morning. I often sit down to eat at my desk but have to come home at night to recover.
GPT-2 CD-GT	It started when I was in school. My father knocked on my door asking if I wanted to come with him to the cafeteria to lunch. When I opened the door, I saw my father lying on the ground with a head wound. My father rushed to the scene to save him, But when I pressed him, I saw his bloody nose cut and swollen chin. My father was bleeding badly from the wound.
GPT-2 CD-pred	It started when I was in school. My father knocked on my door asking if I wanted to come with him to the cafeteria to lunch. When I opened the door, I saw my father lying on the ground with a head wound. My father rushed to the scene to save him, But when I pressed him, I saw his bloody nose cut and swollen chin. My father was bleeding badly from the wound.
GPT-2 Medium	I heard the news about the student strike and decided to go to the university administration. I asked them to give me a rest for a few days. But they didn't listen to me. I was under the impression that they would listen to me. But then I heard the news about the strike again. I went back to the university administration and again tried to get some help.
GPT-2 Medium CD-GT	Well, I just stood up too fast. I feel dizzy.
GPT-2 Medium CD-pred	Well, I just stood up too fast. I feel dizzy.
DialoGPT	
DialoGPT small	It's a little complicated. First, It started out as a fever, And now it's a headache.
DialoGPT small CD-GT	It's really a mystery.
DialoGPT small CD-GT	It's really a mystery.
DialoGPT medium	I guess I had a bad dream last night.
DialoGPT medium CD-GT	Well, I was under the influence when I got up this morning.
DialoGPT medium CD-pred	Well, I was under the influence when I got up this morning.
BART	
BART base	What's the matter with you?
BART base CD-GT	I'm afraid you have a bad headache.
BART base CD-pred	I'm afraid you have a bad headache.
BART large	Well, Young man, I'm afraid you'll have to stay in bed for three days.
BART large CD-GT	I see. Well, I'll prescribe some ointment and you'll be well soon.
BART large CD-GT	I see. Well, I'll prescribe some ointment and you'll be well soon.
Beluga	
Beluga	Yes, it feels like a sharp pain and it's really intense. It's like somebody is hammering inside my head.
Beluga F&R CD-GT	Good morning, what's the matter with you?
Beluga F&R CD-pred	Good morning, what's the matter with you?
Beluga PB CD-GT	Good morning, doctor. I have a terrible headache. The headache started late last night after I spent several hours studying.
Beluga PB CD-pred	What were you doing before you got the headache?

TABLE IV
EXAMPLE OF THE RESPONSE OBTAINED BY ALL THE CONSIDERED APPROACHES TO A SAME CONTEXT SAMPLE.

IV. APPENDIX D: COMPLETE FORM FOR HUMAN EVALUATION STEP 3

In this Appendix, we present the details and reference materials that were provided to the human judges during the evaluation task. Steps 1 and 2 relied on the definitions for Consistency and Specificity given in the paper. For Step 3, the annotators first had to tag each response with dialogue responses. Daily Dialog uses a system of 4 dialogue acts:

For example, the response: I'm sorry to hear about Suzy's cold. Do you think you could ask someone from the family or close friends to help out? It might be best not to take her on the trip if she's not feeling well.

Will be tagged as: <I> I'm sorry to hear about Suzy's cold.</I> <Q> Do you think you could ask someone from the family or close friends to help out?</Q> <I> It might be best not to take her on the trip if she's not feeling well.</I>

Once the response is annotated with the dialogue acts, the judges must rate the following items:

Code	Strategy	Description	Examples
I	**inform**	Provide an information. *Inform, clarify / explain / reply, statement*	I like cooking by myself. I like to taste delicious food.
Q	**question**	Ask for / seek an information. *Ask information, query, open questions, rhetorical questions, repeat question, yes/no-questions, other Qs*	Anyone home? Do you want black or white coffee?
D	**directive**	Directive is a speech where the speaker commands the interlocutor to do something. *Directives: commands, requests, challenges, invitations, orders, summons, entreaties, dares, elicit, offer, or suggest, instruct*	I would like to register for a class today. How about another coffee? Make sure to take proper care of this video. Give me a call and let's go down together.
C	**commissive**	Commissive is a speech where the speaker acts for future action, such as promising or offering. *Commissive: promises, oaths, pledges, threats, vows, offer, acknowledge, commit*	I can show it to you now if you like. I don't wanna be involved in your quarrel.

Fig. 1. 4 dialogue acts used to annotate Daily Dialog, as well as some examples from the dataset to assist this task.

Item	Instruction	Rating Scale
LOGICAL		
Interestingness	Does the response add something to the interaction, is it unexpected, amusing, or insightful? Does it encourage further conversation? Is it engaging in form/content (anecdote etc... not on the topic of discussion) and denotes the speaker's effort to make the conversation engaging.	**1** | The answer is not at all specific, and is not consistent with the context
		2 |The answer makes sense
		3 | The answer is interesting, specific and invites further conversation.
Fluency	Measure the quality of individual sentences. Sentences in a fluent summary should be free of formatting problems, capitalization errors, or obviously ungrammatical sentences (e.g., fragments, missing elements) that make the text difficult to read.	**1** | Not fluent
		2 |Acceptable fluency
		3 | Very fluent
Style	Does the style of the response matches that of the context? Formal / casual, written / oral...	**0** | Style is not matching context
		1 |Style is matching context
Are « social norms » respected?		
EMOTIONAL		
Presence of Emotion	Is an emotion expressed in the answer?	On the evaluation interface, select the emotions expressed in the statement from the choices given (happiness, sadness, anger, disgust, fear, neutral, surprise).
Emotion Adequacy	Does the tone of your answer seem appropriate to the context?	**1** | The emotional tone of the answer doesn't make sense, and / or is confusing
		2 |Emotional tone roughly matches, it's not perfect but it's passable.
		3 | The emotional tone of the response is consistent with the tone and context of the conversation and respects the conversational/social expectations of the interaction.
SOCIAL		
Social Adequacy	The conversation takes place in a specific social context. Are the social conventions of this context respected? Are the dialogue strategies annotated beforehand appropriate (to the context of the conversation and the historical extract)?	**1** | Dialogue strategy(ies) used do not make sense and are confusing
		2 |The dialogue strategies used are not optimal but are viable.
		3 | The dialogue strategies used are coherent and respect the conversational/social expectations of the interaction (answering a question, etc.).
Role Consistency	In a dialogue between 2 speakers, it can happen that a generative model takes the wrong role and answers in speaker A's place, using the wrong pronouns... The aim is to assess whether the response is indeed in the right role. Is the speaker's role respected?	**0** | Not consistent with role (becomes speaker A instead of speaker B, doesn't use correct pronouns, gets the situation wrong and answers for the other speaker...)
		1 |Respects the role he/she plays, uses the right pronouns, and doesn't get the situation wrong

Fig. 2. Definition of each socio-emotional criteria rated in this evaluation, as well as the rating scale used for each item

V. Appendix E: Detailed Results of Human Evaluation

In Table V, you will find the details of all scores obtained from the human evaluation we carried out on Daily Dialog. While the *socemo* score is weighted by the number of responses by the model in the annotated sample, the logical, emotional and social ratings are unweighted. We weigh the fluency score similarly to the *socemo* score to compare it to the Perplexity metric.

VI. Appendix F: Results on New Daily Dialog Dataset

Instead of using the huggingface dataset, which was reported to have a significant overlap between the test and train sets, we use the splits provided in Daily Dialog's original paper, which do not display the same duplicate issue. In our original experiments, we had not fine-tuned our Beluga models (inference only), so those results are unaffected by the test-train set data overlap. We reran our code on the remaining models - BART, DialoGPT and GPT2 - using the same GPU and hyper-parameters as in the main paper). These results, available in Table VI are similar to those obtained with the test-train sets duplicates. While we do not claim that using the huggingface splits displaying duplicates did not have any negative impact on the training, this new set of results seems to indicate that this impact might not be too significant or invalidate the results shown in this study.

Model	filtered	top3	socemo	logical	emotional	social	weighted fluency
GPT-2							
GPT-2 Small NO-CD	33	9	13	90	**100**	94	12
GPT-2 Small CD-pred	37	9	14	88	98	98	13
GPT-2 Small CD-GT	37	9	14	88	98	98	13
GPT-2 Medium NO-CD	**53**	21	28	91	99	98	27
GPT-2 Medium CD-pred	**5**	19	30	83	99	99	29
GPT-2 Medium CD-GT	**5**	19	30	93	99	99	29
DialoGPT							
DialoGPT Small NO-CD	37	11	18	90	98	98	17
DialoGPT Small CD-pred	4	13	19	90	97	98	18
DialoGPT Small CD-GT	4	13	19	90	97	98	18
DialoGPT Medium NO-CD	**53**	16	15	89	99	99	14
DialoGPT Medium CD-pred	**52**	16	20	88	**100**	**100**	18
DialoGPT Medium CD-GT	**52**	16	20	88	**100**	**100**	18
BART							
BART Base NO-CD	32	7	13	87	97	95	12
BART Base CD-pred	32	7	14	82	98	96	13
BART Base CD-GT	32	7	14	82	98	96	13
BART Large NO-CD	42	9	19	89	99	99	19
BART Large CD-pred	45	12	19	88	**100**	99	19
BART Large CD-GT	45	12	19	88	**100**	99	19
Beluga							
Beluga NO-CD	42	25	39	93	98	**100**	38
Beluga PB CD-pred	**51**	**36**	44	93	98	98	44
Beluga PB CD-GT	45	3	**51**	94	97	99	52
Daily Dialog Reference	97	61	69	94	98	**100**	63

TABLE V

ALL THE RESULTS FROM THE HUMAN EVALUATION: STEP 1 - FILTERING (COLUMN 1), STEP 2 - TOP-3 (COLUMN 2) & STEP 3 SOCIO-EMOTIONAL ANNOTATION (COLUMN 3 IS THE GLOBAL SCORE, COMPUTED AS THE AVERAGE OF THE THREE AXES SCORES IN COLUMNS 4-6).

Model	Sacre bleu	Rouge	Bert score	CHRF
GPT-2 $_b$ NO-CD	96	12	86	13
GPT-2 $_b$ CD-pred (F&R)	103	12	86	15
GPT-2 $_b$ CD-GT (F&R)	103	12	86	15
GPT-2 $_M$ NO-CD	176	14	87	14
GPT-2 $_M$ CD-pred (F&R)	169	14	87	16
GPT-2 $_M$ CD-GT (F&R)	169	14	87	16
DialoGPT $_b$ NO-CD	99	13	86	12
DialoGPT $_b$ CD-pred (F&R)	90	13	87	15
DialoGPT $_b$ CD-GT (F&R)	90	13	87	15
DialoGPT $_M$ NO-CD	217	15	87	14
DialoGPT $_M$ CD-pred (F&R)	233	16	87	17
DialoGPT $_M$ CD-GT (F&R)	233	16	87	17
BART $_b$ NO-CD	218	17	88	12
BART $_b$ CD-pred (F&R)	236	17	87	18
BART $_b$ CD-GT (F&R)	236	17	87	18
BART $_L$ NO-CD	303	18	87	14
BART $_L$ CD-pred (F&R)	356	19	87	20
BART $_L$ CD-GT (F&R)	236	16	87	18

TABLE VI

COMPARATIVE RESULTS OF THE EXPERIMENTS ON CONDITIONING RESPONSE GENERATION USING MULTI-LABEL SEQUENCES MODELLING SOCIAL AND EMOTIONAL BEHAVIOURS, ON A DIFFERENT DAILYDIALOG SPLIT, THAT DOES NOT FEATURE ANY DUPLICATE ACROSS THE DIFFERENT SETS. RESULTS ARE GIVEN IN %. $_B$ DENOTES THE BASE OR SMALL MODEL, $_M$ THE MEDIUM MODEL AND $_L$ THE LARGE MODEL.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019.

[3] 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.