Supplementary materials for Mimic and Rephrase: Generating empathetic responses in open-ended dialogue

A Appendix

In this appendix, we provide additional examples, and error cases as well as screenshots of the interfaces we used for conducting human judgements of the evaluation metrics, and correlation between the human evaluation metrics.

Table 1 shows additional examples from the IDONTKNOW and EMOTIVE datasets along with the human assessment of **Goodness**, **Appropriateness**, **Fluency**, and **Specificity** as well as the automatic metrics of **Conciseness** and **Repetitiveness**.

Figure 1 shows the distribution of how often rules are triggered for the IDONTKNOW dataset.

Table 2 shows examples where the BLEU and METEOR scores are not reflective of the quality of the response, motivating the need for collecting human assessment.

Table 3 provides a breakdown of classes of errors that were observed.

Table 4 provides additional examples for the **S2SA+C** model.

Figures 2, 3, 4 illustrates the interfaces and instructions we presented Amazon Mechanical Turkers for soliciting human assessment of **Appropriateness**, **Specificity**, and **Fluency**.

Figures 5, 6 shows the correlation between different human evaluation metrics for the IDONTKNOW and EMOTIVE responses.

	IDONTKNOW	Good	App	Flu	Spec	Con	Rep
a	**P** So this customer is trying to purchase this lamp but it does n't have a UPC code on it , do I go to the back and find another one just like it to scan **MR** i do not know if you should find another one or not	3.2	0.33	0.83	0.67	0.34	0.14
b	**P** I lost my pass and need a replacement . What should I do **MR** i do not know how you should get a replacement pass	3.6	1	1	1	0.91	0.40
c	**P** Can I take photographs in the museum using , for example , a digital camera **MR** i do not know if you can photographs using a digital camera	3.8	0.33	0.5	0.67	0.79	0.39
d	**P** My child got on my computer and bought something that I do n't want , I need to cancel this order **MR** i am not able to help you cancel the order	4.4	1	1	1	0.44	0.10
e	**P** I want this invoice sent to my fax **MR** i am not able to send the invoice to your fax	4.4	1	1	0.33	1.31	0.24
f	**P** How long is the audition class **MR** i do not know how long the audition class is	4.8	1	1	1	1.45	0.88
	EMOTIVE	Good	App	Flu	Spec	Con	Rep
a	**P** I got mugged walking home from work Friday **MR** Sorry I got mugged walking home from work Friday	1.4	1	0.5	0.67	1.13	0.99
b	**P** I accidentally put a $ 100 into the Salvation Army donation bucket instead of a $ 1 , and that was my grocery money **MR** I am sorry that you accidentally put a $ 100 into the Salvation Army donation bucket instead of a $ 1 and it was your grocery money	3.2	1	1	1	1.13	0.90
c	**P** The new socks my mom gave me for Christmas already have a hole in them **MR** I 'm sad that your new socks from your mom already have holes in them	3.8	1	0.83	1	0.99	0.59
d	**P** The plane was rerouted twice and we lost all of our luggage **MR** I am sorry you lost your luggage	4	0.33	1	0	0.56	0.20
e	**P** The fencing for the garden was completed on time **MR** Good to hear that the fencing of the garden was completed on time	4	1	1	1	1.29	0.77
f	**P** I proposed to my girlfriend this morning and she accepted **MR** I am happy to hear you are engaged	4	1	1	0	0.62	0.14
g	**P** My team won the game ! **MR** Super , that is incredible news	4.6	1	0.67	0	1.47	0
h	**P** Kim does n't feel the same way about me as I do about her **MR** thats unfortunate , i always thought you and Kim had reciprocal feelings for each other	4.8	1	1	0.67	1.47	0.13

Table 1: Examples from the IDONTKNOW and EMOTIVE dataset with corresponding metrics of **Good**ness, **App**ropriateness, **Flu**ency, **Spec**ificity, **Con**ciseness, and **Rep**etitiveness. Note that EMOTIVE examples a is poor since the pronoun is incorrect, and b is overly repetitive. While it is okay to repeat parts of the prompt exactly, the best responses repeat only the relevants parts of the prompt (see IDONTKNOW d) or capturing the key parts of the prompt using different words (see EMOTIVE f,h). In all cases, a good response typically must be judged to be appropriate and fluent.

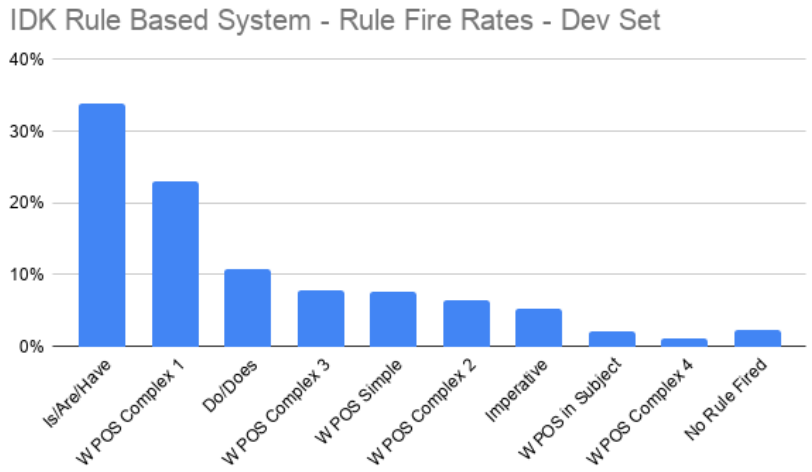

Figure 1: Distribution of how often each rule is triggered in the rule based system for the IDONTKNOW dataset (computed on the dev set).

	IDONTKNOW	EMOTIVE
Prompt	Do you all have this dress in a size 6	My fridge is fixed finally
Human	I do not know if we have this dress in a size 6	I am glad to hear about your fridge
S2S+attn	i do not know if we have this dress in a size	i am glad to hear about your fish
BLEU	92.0	84.1
METEOR	93.0	87.4
Prompt	I bought a couple of things right now , checked out , but realized I forgot a couple of things and am wondering if I can add them to my order	Wifi has been down at work for the past week
Human	i am not able to add the things that you forgot to your order	i am sorry wifi has been down
S2S+attn+copy	i do not know if you can add them to your order	i am sorry to hear about your wifi
BLEU	0	0
METEOR	41.5	52.8

Table 2: Examples where the BLEU and METEOR scores are not a good indications of quality. Top: High BLEU and METEOR scores, but inappropriate responses. Bottom: Low BLEU and METEOR score, but good responses.

Error	Model	Example
Not appropriate		
Off topic	S2SA	**P** Does the Subaru Certified Pre-Owned program have a deductible **H** I do not know if the Subaru Certified Pre-Owned programs has a deductible **S2SA** i do not know if the drone federal reserve program program has a tax
Missing key term	S2SA+C	**P** My daughter brought home an excellent report card today **H** I am happy about your daughters report card **S2SA+C** i am glad to hear about your report
Pronoun Error	Rule-IDONTKNOW	**P** Do you guys offer repair services for items **H** I do not know if we offer repair services for items **RB** I do not know if I guys offer repair services for items
Not fluent		
Punctuation	S2SA	**P** What constitutes a ' public ' screening **H** I do not know what constitutes a public screening **S2SA** i do not know what ' a ' ' ' '
Repeating Words	S2SA	**P** I wanted to buy a game but it was already sold out **H** I am sorry the game was sold out **S2SA** i am sorry to hear about the game game
Verb Coordination	Rule-IDONTKNOW	**P** Did John Wayne shoot any films in Lone Pine **H** I do not know if John Wayne shot any films in Lone Pine **RB** i do not know if john wayne shoot any films in lone pine
Incorrect reordering	Rule-IDONTKNOW	**P** What if the database does not contain the name of the person I am looking for **H** i do not know if john wayne shot any films in lone pine **RB** I do not know what the database if not contains the name of the person you are looking for

Table 3: Breakdown of cases where the response is inappropriate or not fluent. The models (**Rule**-based, **S2SA**, **S2SA+C**) in which the type of error occurs is listed in column 2.

Metric	Model	Example
Low repetitiveness	S2SA+C	IDONTKNOW **P** We are a family group of 5 wanting entrance tickets to the Vatican . 3 are over 18 and 2 under . It is unclear whether I should book for the under 18 separately as although I mentioned all 5 names in the group space , the quoted price seems to be only for the over 18 . Pls advise **H** i do not know if you should book for the under 18 separately **S2SA+C** i am not able to advise you
High repetitiveness	S2SA+C	IDONTKNOW **P** Should an employer seek a refund from the IRS of overwithheld income taxes due to a prior-year wage overpayment **H** i do not know if an employer can make up for a wage overpayment by paying the employee less in the next wage payment **S2SA+C** i do not know if an employer should seek a refund from the irs of overwithheld income taxes due to a prior-year wage overpayment
Low Specificity	S2SA+C	EMOTIVE **P** I was recommended for a new position **H** I am glad you were recommended **S2SA+C** i am glad to hear about your position

Table 4: Examples where the responses for the **S2SA+C** model is good for both low and high repetitiveness, and low specificity.

General instructions

- You will be presented with several examples where Blue (a machine) is responding to Alex (a person) and must **rate the appropriateness of Blue's responses** according to one of the following prompts:
  - "I don't know...": Did Blue appropriately convey that it either doesn't know the answer to Alex's question or is unable to help?
  - "I am glad...": Did Blue appropriately empathize with Alex's happy story?
  - "I am sorry..": Did Blue appropriately empathize with Alex's sad story?
- Each task will have several examples: **consider every example to be from a separate conversation**.
- You will be able to submit once all of Blue's responses have been rated.

Tutorial

Answer each of the following examples to proceed.

Example (I am glad...)

That's right! While generic, Blue has conveyed that it is happy for Alex.

Sentence	Appropriateness
Alex: I found a dollar. **Blue:** I am happy you found a dollar.	✔ ✖

Example (I am sorry...)

That's right! Blue did not understand why Alex was sad, and should have said "I'm sorry your mom lost her wallet" instead.

Sentence	Appropriateness
Alex: My mom lost her wallet in the store. **Blue:** I'm sorry your mom was in the store.	✔ ✖

Example (I don't know...)

That's right! Blue's response clearly didn't address what Alex was asking and should have responded with "I don't know if you need to bring a passport" instead.

Sentence	Appropriateness
Alex: If I'm flying to Hawaii, do I need to bring a passport? **Blue:** I don't know if you are flying to Hawaii.	✔ ✖

Figure 2: Instructions for interface used to solicit human assessment of appropriateness

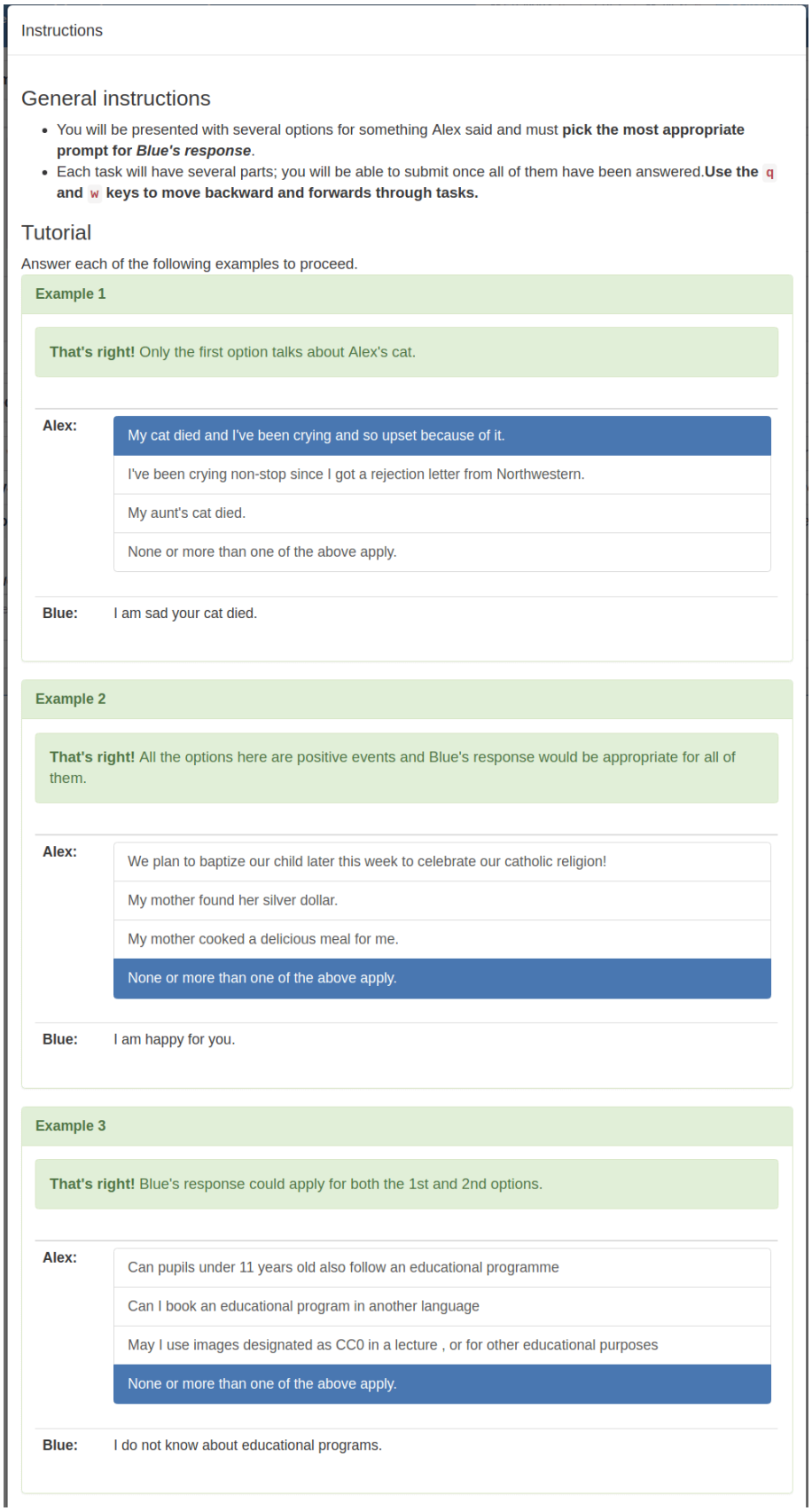

Figure 3: Instructions for interface used to solicit human assessment of specificity

General instructions

- You will be presented with several sentences that were generated by a machine and must rate the fluency of the responses.
- A fluent sentence should be **both grammatical and make sense**: *The president hoped that that would not happen.* is grammatical while *I hope that that I hope you like the color* is not.
- Each task will have several sentences; you will be able to submit once all of them have been rated.

Tutorial

Answer each of the following examples to proceed.

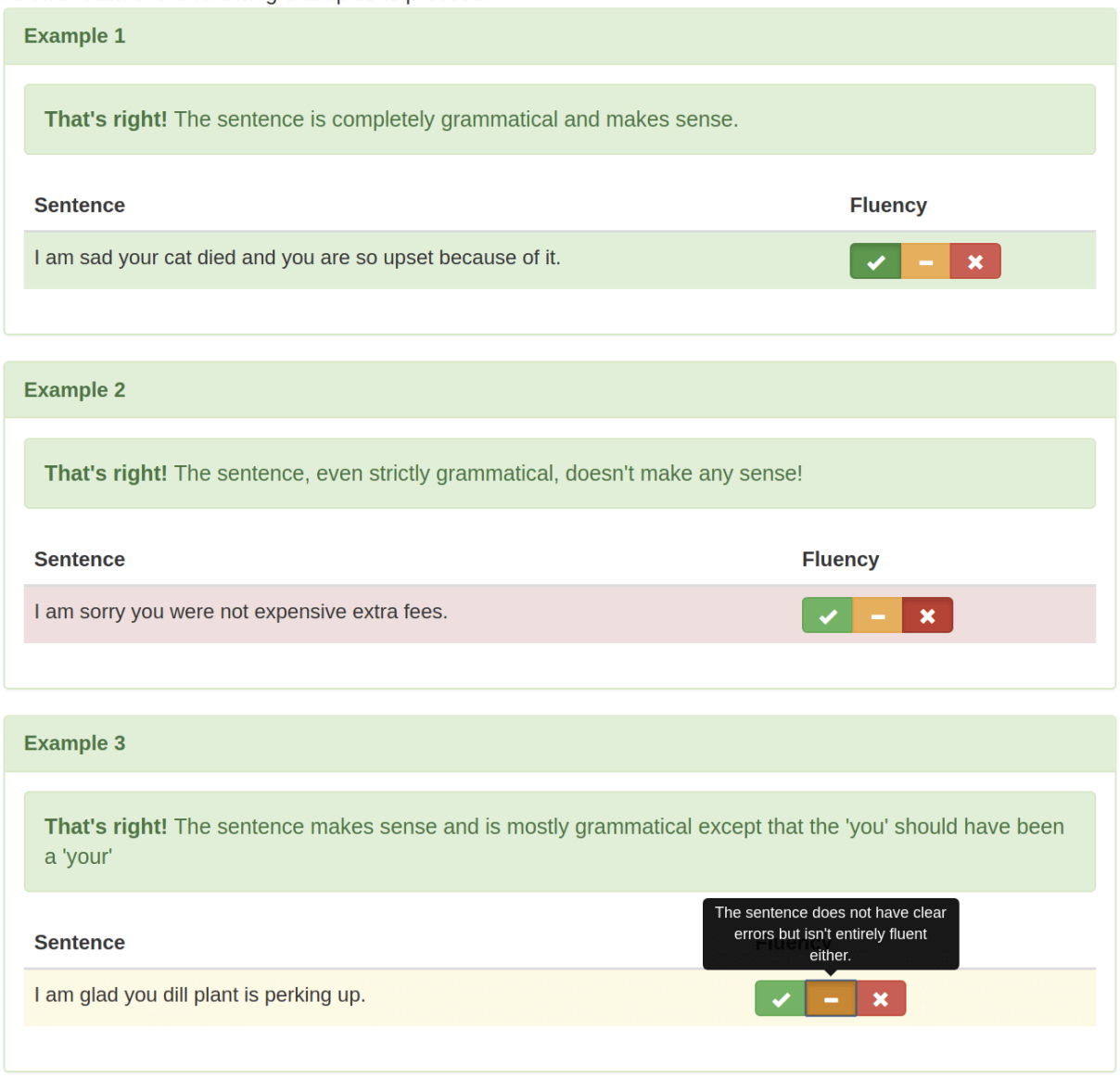

Figure 4: Instructions for interface used to solicit human assessment of fluency

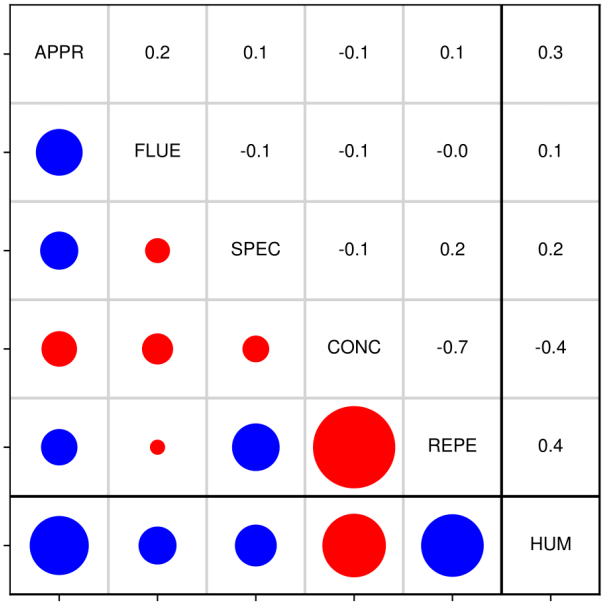

Figure 5: Correlation of human evaluated metrics for IDONTKNOW responses in the dataset. We include the evaluations of appropriateness, fluency, specificity, conciseness, and repetitiveness. In the diagram, HUM refers to the Likert scale evaluations of overall response quality.

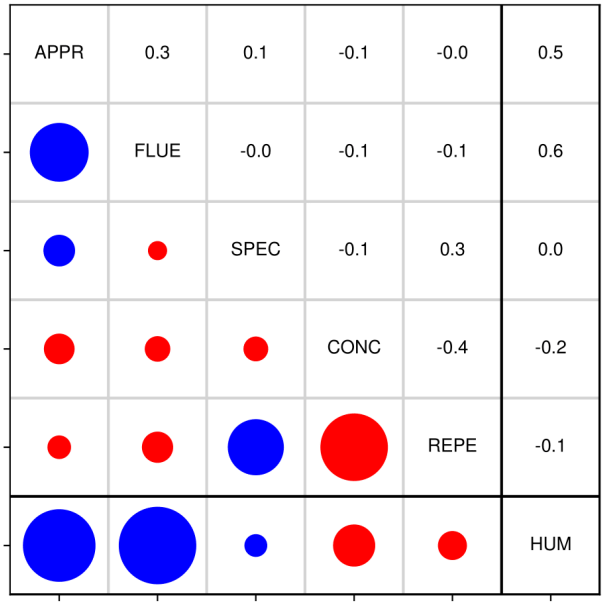

Figure 6: Correlation of human evaluated metrics for EMOTIVE responses. We include the evaluations of appropriateness, fluency, specificity, conciseness, and repetitiveness. In the diagram, HUM refers to the Likert scale evaluations of overall response quality.