A Data Examples

We include ten randomly selected dialogues from our training set in Table 5.

B Human Evaluation Crowdsourcing Task

Human evaluations were collected on MTurk. For the rating task, each worker was shown a set of 10 randomly subsampled examples from the test set, one after another, each from a different randomly selected model. The worker had to rate the empathy, relevance, and fluency of each example before moving onto the next one. At least 100 ratings were collected per model. 221 US workers participated in the rating task, and each had to perform a minimum of one set of 10 ratings.

C Next utterance prediction on other datasets

We test how fine-tuning on ED data affects next utterance prediction on two external datasets (REDDIT and DAILYDIALOG). In this experiment, we use both candidates and context from the DD or R data. Results in Table 6 show that performance on DAILYDIALOG improves after fine-tuning on our data.

D Additional Experimental Details and Results

D.1 Training Details

We used Adamax for training throughout, and dropout was set to 0% everywhere except for a 20% dropout in the linear layer of the emotion-label term of the MULTITASK objective function (discussed below). A learning rate of $8\mathrm{e}{-4}$ was used for all four-layer Transformer models, following Mazare et al. (2018). For the five-layer retrieval-based Transformer model (Pretrained-Large and Fine-Tuned-Large), the learning rate was selected by picking the best performing over the validation set, among values randomly sampled between between $5\mathrm{e}{-5}$ and $8\mathrm{e}{-4}$. When training the retrieval-based BERT model on Reddit and ED data, the learning rate was selected by picking the best performing over the validation set, among values randomly sampled between $6\mathrm{e}{-6}$ and $2\mathrm{e}{-4}$. For training BERT models on Reddit data, we also experimented with adding an additional Transformer layer after the output embedding of the BERT model, but this slightly degraded P@1,100 scores on the validation set. We used a learning rate of $8\mathrm{e}{-5}$ for the five-layer generative Transformer models.

D.2 Additional Experimental Set-Ups

We investigated a few additional approaches for incorporating supervised emotion or topic prediction in generating dialogue, but observed little performance improvement. Methods are described below.

Multitask with Emotion labels If the most appropriate response depends on some information for which supervision is available, e.g., the emotions at play, nudging the model to encode this information could result in better performance. We experimented with this by training the base architecture in the one-to-many style of multi-task learning that has been used for NLP seq2seq settings (Luong et al., 2016). In this set-up, MULTITASK, we altered the objective function to also optimize for predicting the emotion label of the conversation to which the utterances being encoded belonged. We added to the context encoder a linear layer and softmax that predicted the emotion label from the context sentences. The objective function was altered to be the average of the negative log-likelihood of predicting the next utterance $\bar{y}$ and the negative log-likelihood of the added linear layer being able to predict the correct emotion.

Prepend-3, Prepend-5 We investigated whether Prepend models could be improved by adding the top-3/5 predicted emotion or topic labels by the classifier (rather than top-1).

Ensemble of Encoders We also investigated another approach for incorporating external predictors, which we report the results of in our extended results tables. In this set-up (ENSEM), we augmented the encoders to incorporate latent representations from pretrained supervised architectures. We replaced each of the encoders in our Transformer networks with an Ensemble encoder, similar to a many-to-one style encoder-decoder architecture (Luong et al., 2016). This encoder took the encoding h_w from our basic Transformer encoder (either h_x or h_y), already trained on our data, and concatenated it with the representation h_c extracted from the inner layer of a classification network. We used the penultimate layer of a deep emotion classifier. The concatenated encodings were projected linearly to the dimension required

Label: Content
Situation: Speaker felt this when...
"eating my favorite meal makes me happy."
Conversation:
Speaker: i am at my best when i have my favorite meal.
Listener: nice
Speaker: i love enchiladas
Listener: really?
Speaker: yes. enchiladas for the win!

Label: Proud
Situation: Speaker felt this when...
"I was proud when my brother finished college. He worked so hard at it"
Conversation:
Speaker: I was proud of my brother when he finished school. He worked so hard at it
Listener: Nice, tell him congrats. What did he major in?
Speaker: It was English
Listener: He should become an English teacher1

Label: Joyful
Situation: Speaker felt this when...
"I have had a great week!"
Conversation:
Speaker: I have had a great start to my week!
Listener: That's great. Do you think the rest of the week will be as great?
Speaker: I hope so! It looks promising!!
Listener: Lucky you. Are you always a positive person or it's just been an amazing week really?
Speaker: haha. Kind of both. And also probably too much coffee to start my shift tonight

Label: Terrified
Situation: Speaker felt this when...
"I got home for lunch and found a bat outside on my front porch."
Conversation:
Speaker: I got home for lunch and found a bat outside on my front porch. It probably has rabies. Bats shouldn't be out during the day.
Listener: Doesn't rabies cause sensativity to light? Either way I would freak out...
Speaker: It can but, it also causes anmails to behave erratically... like bats wadering around in the middle of the day.
Listener: Oh yeah, gotcha. I really don't like animals that are small and move quickly
Speaker: Generally yes.

Label: Anticipating
Situation: Speaker felt this when...
"I cant wait to go on my end of summer trip"
Conversation:
Speaker: I cant wait to go on my end of summer trip in texas.
Listener: Sounds like fun. What you got planned ?
Speaker: not really sure but im excited to just be invited
Listener: Got any family out there? Cousins perhaps

Label: Terrified
Situation: Speaker felt this when...
"My brother jump scared me while I was out playing. It was crazy bad."
Conversation:
Speaker: Just got scared to death.
Listener: Oh no. What happened?
Speaker: My brother jumped scared me.
Listener: lol is he younger or older?

Label: Proud
Situation: Speaker felt this when...
"My little dog learned to sit!"
Conversation:
Speaker: I finally tough my new little puppy his first trick!
Listener: What trick did you teach him?
Speaker: I tought him to sit for a treat, its so cute.
Listener: That is good, do you plan to teach him more tricks?

Label: Apprehensive
Situation: Speaker felt this when...
"I have to call my landlord about being late on the rent. I really don't want to have this conversation."
Conversation:
Speaker: I have to make a dreadful phone call tomorrow
Listener: Oh no, about what?
Speaker: I'm late on my rent and I need another week. I don't want to because my landlord isnt very nice
Listener: Oh no, I've been there done that too many times.
Speaker: I don't want her to make a big deal

Label: Confident
Situation: Speaker felt this when...
"When my husband asked me about how to build a chicken coop I was able to give him a reply that was backed up by blueprints and research from the internet. "
Conversation:
Speaker: We recently got 9 chicks and we've been having to work on making them a coop! I had to do so much research but I think we finally have a place that they'll enjoy living when they aren't able to free range.
Listener: OHH! I Love chickens ! I have always wanted some. I have a duck! lol- What kind of chickens are they?
Speaker: We currently have 2 Australorps, 3 Rhode Island Reds, 3 Barred Plymouth Rocks, and 1 Welsummer, but 4 of the 9 ended up being roosters. Ugh!
Listener: Oh man! They fight sometimes. I hope they aren't too bad about waking you up in the morning. Chickens can be very sweet though!
Speaker: I love my little hens, especially one I've named Curly. The roosters might get replaced by hens though because the crowing is so frustrating!

Label: Surprised
Situation: Speaker felt this when...
"I got a lottery ticket while I was at work today. I won $100 on the scratch off. I was shocked. I never win."
Conversation:
Speaker: I won $100 on a scratch off today. I was shocked. I never win.
Listener: Wow! How often do you play the lottery?
Speaker: I usually go on our Tuesday break to buy one with coworkers.
Listener: Neat! Well that is a fantastic feat. Maybe you can win again sometime?

Table 5: 10 random examples from EMPATHETICDIALOGUES training set.

	P @1,100		BLEU	
Model	DD	R	DD	R
Pretrained	39.04	58.95	6.65	1.43
Fine-Tuned	44.58	56.25	7.14	1.64
Pretrained-Large	42.28	61.60	6.94	1.42
Fine-Tuned-Large	48.96	58.71	7.42	1.73

Table 6: Performance of the retrieval-based pretrained model and retrieval-based models fine-tuned on ED data for next utterance prediction in other datasets, with both context and candidates from the same dataset (R=Reddit, DD=DailyDialog).

by the decoder, whose architecture didn't change. When training the dialogue model, we froze both the base Transformer encoder and the pretrained classifier and trained only the linear layers (and the decoder for generative systems). We used emotion-related supervision from Emojis from Twitter, through the use of the trained Deepmoji system (Felbo et al., 2017) released by the authors, either as-is (ENSEM-DM) or fine-tuned on the situation descriptions of EMPATHETICDIALOGUES (ENSEM-DM+).

D.3 Additional Experiments Results

Automated and human evaluations for any additional experiments are in Tables 7 and 8, respectively. All of these model variations show improvements over the pre-trained models. In some metrics, many of these models show slight improvements over the fine-tuned models, as well, though not as consistently, except for the larger BERT retrieval-based models. While prepending top-1 or top-3 labels do not improve generative model scores, the results in Table 8 suggest that multitask, prepend-5, and ensemble set-ups may improve the human evaluations of the fine-tuned generative model for empathy, but are too inconsistent to be conclusive without more corroborating experiments.

D.4 Emotion Classification Results

Our dataset can also be used to train or fine-tune an emotion classifier, as we do in our PREPEND-K and ENSEM-DM+ set-ups. To give a sense of where the difficulty falls compared to existing emotion and sentiment classification benchmarks, we reproduce the table from Felbo et al. (2017) and add results when fine-tuning the Deepmoji model on our dataset, or using a fastText classifier (Table 10).

		Retrieval		Retrieval w/ BERT		Generative	
Model	Candidate Source	P@1,100	AVG BLEU	P@1,100	AVG BLEU	PPL	AVG BLEU
Pretrained	R	-	4.10	-	4.26	27.96	5.01
	R+ED	-	4.96	-	5.62	-	-
	ED	43.25	5.51	49.94	5.97	-	-
Fine-Tuned	R	-	3.85	-	4.14	-	-
	R+ED	-	4.76	-	5.39	-	-
	ED	56.90	5.88	**65.92**	6.21	21.24	6.27
	ED+DD	-	5.61	-	-	-	-
	ED+DD+R	-	4.74	-	-	-	-
Pretrained-Large	R	-	4.16	-	-	-	-
	ED	47.58	5.78	-	-	23.64	6.31
Fine-Tuned-Large	ED	**60.44**	6.01	-	-	**16.55**	**8.06**
Multitask	ED	55.73	6.18	65.90	6.17	24.07	5.42
EmoPrepend-1	ED	56.31	5.93	66.04	6.20	24.30	4.36
EmoPrepend-3	ED	55.75	**6.23**	65.85	6.14	23.96	2.69
EmoPrepend-5	ED	56.35	6.18	64.69	6.21	25.40	5.56
TopicPrepend-1	ED	56.38	6.00	65.96	6.18	25.40	4.17
TopicPrepend-3	ED	55.44	5.97	65.85	**6.25**	25.02	3.13
TopicPrepend-5	ED	55.75	6.17	65.65	6.19	25.10	6.20
Ensem-DM	ED	52.71	6.03	-	-	19.05	6.83
Ensem-DM+	ED	52.35	6.04	-	-	19.10	6.77

Table 7: Automatic evaluation metrics on the test set for full set of experimental setups. Pretrained: basic Transformer model pretrained on a dump of 1.7 billion Reddit conversations. Fine-Tuned: model fine-tuned over the EmpatheticDialogues training data. Multitask: model trained with multitask loss function (predicting the emotion label). EmoPrepend-1/3/5, TopicPrepend-1/3/5: model using top-k labels outputted by an external classifier as prepended tokens. Ensem: model incorporating external classifiers by concatenating representations from deepmoji with the fine-tuned transformer representation. Candidates come from Reddit (R) or EmpatheticDialogues (ED). P@1,100: precision retrieving the correct test candidate out of 100 test candidates. AVG BLEU: average of BLEU-1,-2,-3,-4. PPL: perplexity. *Bold: best performance in that column.*

	Model	Candidate	Empathy	Relevance	Fluency
Retrieval	*Pretrained*	R	2.82 ± 0.12	3.03 ± 0.13	4.14 ± 0.10
		R+ED	3.16 ± 0.14	3.35 ± 0.13	4.16 ± 0.11
		ED	3.45 ± 0.12	3.55 ± 0.13	4.47 ± 0.08
	Fine-Tuned	R	2.51 ± 0.12	2.90 ± 0.12	4.04 ± 0.11
		R+ED	3.06 ± 0.14	3.34 ± 0.13	4.12 ± 0.11
		ED	3.76 ± 0.11	3.76 ± 0.12	4.37 ± 0.09
	Multitask	ED	3.63 ± 0.12	3.83 ± 0.12	4.49 ± 0.08
	EmoPrepend-1	ED	3.44 ± 0.11	3.70 ± 0.11	4.40 ± 0.08
	EmoPrepend-3	ED	3.54 ± 0.11	3.76 ± 0.11	4.54 ± 0.07
	EmoPrepend-5	ED	3.42 ± 0.11	3.61 ± 0.11	4.53 ± 0.07
	TopicPrepend-1	ED	3.72 ± 0.12	3.91 ± 0.11	4.57 ± 0.07
	TopicPrepend-3	ED	3.64 ± 0.11	3.66 ± 0.12	4.51 ± 0.08
	TopicPrepend-5	ED	3.34 ± 0.12	3.52 ± 0.12	4.24 ± 0.09
	Ensem-DM	ED	3.61 ± 0.11	3.71 ± 0.12	4.45 ± 0.08
	Pretrained-Large	R	2.94 ± 0.14	3.12 ± 0.14	4.23 ± 0.10
		ED	3.47 ± 0.14	3.56 ± 0.13	4.41 ± 0.10
	Fine-Tuned-Large	ED	3.81 ± 0.12	3.90 ± 0.12	4.56 ± 0.08
Retrieval w/ BERT	*Pretrained*	R	3.06 ± 0.13	3.29 ± 0.13	4.20 ± 0.10
		R+ED	3.49 ± 0.12	3.62 ± 0.12	4.41 ± 0.09
		ED	3.43 ± 0.13	3.49 ± 0.14	4.37 ± 0.10
	Fine-Tuned	R	2.90 ± 0.13	3.39 ± 0.13	4.36 ± 0.09
		R+ED	3.46 ± 0.13	3.90 ± 0.12	4.46 ± 0.08
		ED	3.71 ± 0.12	3.76 ± 0.12	4.58 ± 0.06
	Multitask	ED	3.80 ± 0.12	3.97 ± 0.11	4.63 ± 0.07
	EmoPrepend-1	ED	3.93 ± 0.12	3.96 ± 0.13	4.54 ± 0.09
	EmoPrepend-3	ED	3.73 ± 0.13	3.88 ± 0.14	4.60 ± 0.09
	EmoPrepend-5	ED	4.08 ± 0.10	4.10 ± 0.11	4.67 ± 0.07
	TopicPrepend-1	ED	4.03 ± 0.10	3.98 ± 0.11	4.65 ± 0.07
	TopicPrepend-3	ED	3.73 ± 0.12	3.84 ± 0.13	4.52 ± 0.08
	TopicPrepend-5	ED	3.72 ± 0.12	3.80 ± 0.12	4.46 ± 0.09
Generative	*Pretrained*	-	2.31 ± 0.12	2.21 ± 0.11	3.89 ± 0.12
	Fine-Tuned	-	3.25 ± 0.12	3.33 ± 0.12	4.30 ± 0.09
	Multitask	-	3.36 ± 0.13	3.34 ± 0.13	4.21 ± 0.10
	EmoPrepend-1	-	3.16 ± 0.12	3.19 ± 0.13	4.36 ± 0.09
	EmoPrepend-3	-	3.09 ± 0.13	3.02 ± 0.13	4.39 ± 0.09
	EmoPrepend-5	-	3.32 ± 0.12	3.23 ± 0.12	4.35 ± 0.09
	TopicPrepend-1	-	3.09 ± 0.13	3.12 ± 0.13	4.41 ± 0.08
	TopicPrepend-3	-	3.09 ± 0.12	3.34 ± 0.13	4.53 ± 0.08
	TopicPrepend-5	-	3.46 ± 0.13	3.68 ± 0.13	4.60 ± 0.08
	Ensem-DM	-	3.42 ± 0.12	3.45 ± 0.12	4.67 ± 0.06
	Pretrained-Large	-	2.84 ± 0.13	2.97 ± 0.12	4.01 ± 0.11
	Fine-Tuned-Large	-	3.61 ± 0.13	3.62 ± 0.13	4.46 ± 0.10
Gold Response	–	–	4.19 ± 0.10	4.55 ± 0.07	4.68 ± 0.06

Table 8: Human evaluation metrics from rating task for additional experiments.

	Model	Params, resources, train examples	Emp	Rel	Fluent
Retrieval	Pretrained-R	84.3M, 2.5 days, 8 GPUs, 1.7B	2.8	3.0	4.1
	Pretrained-ED	same , same, +22.3k	3.5	3.6	4.5
	Fine-Tuned	same , + 0.5 hour, 1 GPU, +22.3k	3.8	3.8	4.4
	Multitask	+9.6k, + 0.5 hour, 1 GPU, +22.3k	3.6	3.6	4.5
	Pretrained-Large-R	86.5M, 10.5 days, 8 GPUs , 1.7B	2.9	3.1	4.2
	Pretrained-Large-ED	same, same, +22.3k	3.5	3.6	4.4
	Fine-Tuned-Large	same, +1 hour, 1GPU, +22.3k	3.8	3.9	4.6
	Pretrained-BERT-R	217M, 13.5 days, 8 GPUs , 1.7B	3.1	3.3	4.2
	Pretrained-BERT-ED	same, same, +22.3k	3.4	3.5	4.4
	Fine-Tuned-BERT	same, +1 hour, 8 GPUs, +22.3k	3.7	3.8	4.6
	Multitask-BERT	+9.6k, +0.5 hour, 8 GPUs, +22.3k	3.8	4.0	4.6
Generative	Pretrained	85.1M, 2 days, 32 GPUs, 1.7B	2.3	2.2	3.9
	Fine-Tuned	same , +1 hour, 1 GPU, +22.3k	3.3	3.3	4.3
	Multitask	+9.6k, +1 hour, 1 GPU, +22.3k	3.2	3.2	4.3
	Pretrained-Large	86.2M, 2.5 days, 32 GPUs, 1.7B	2.8	3.0	4.0
	Fine-Tuned-Large	same , +0.5 hour, 1 GPU, +22.3k	3.6	3.6	4.5

Table 9: Training resources for different models, with human ratings for empathy (Emp), relevance (Rel) and fluency (Fluent) for full set of experiments. Retrieval-based models use reply candidates from the ED training set (ED) or from Reddit (R). Resource comparisons are relative to the first row of each group. Fine-tuning on ED improves all scores (except for Fluency in one case) while requiring minimal additional training resources. SEM is approximately 0.1

Dataset	Metric	SOTA (in 2017)	fastText	DeepMoji new	DeepMoji full	DeepMoji last	DeepMoji chain-thaw
SE0714	F1	0.34	0.16	0.21	0.31	0.36	0.37
Olympic	F1	0.50	0.38	0.43	0.50	0.61	0.61
PsychExp	F1	0.45	0.44	0.32	0.42	0.56	0.57
SS-Twitter	Acc	0.82	0.68	0.62	0.85	0.87	0.88
SS-Youtube	Acc	0.86	0.75	0.75	0.88	0.92	0.93
SE0614	Acc	0.51	-	0.51	0.54	0.58	0.58
SCv1	F1	0.63	0.60	0.67	0.65	0.68	0.69
SCv2-GEN	F1	0.72	0.69	0.71	0.71	0.74	0.75
ED	Acc	-	0.43	0.40	0.46	0.46	0.48
ED-cut	Acc	-	0.41	0.36	0.42	0.44	0.45

Table 10: Classification performance on EmpatheticDialogues, with the benchmarks proposed in (Felbo et al., 2017) for reference. ED: performance on predicting the emotion label from the situation description. ED-CUT: same, but after having removed all the situation descriptions where the target label was present.