

AUGESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation

Chujie Zheng Sahand Sabour Jiaxin Wen Zheng Zhang Minlie Huang*

The CoAI Group, DCST, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
chujiezhengchn@gmail.com aihuang@tsinghua.edu.cn

Abstract

Crowdsourced dialogue corpora are usually limited in scale and topic coverage due to the expensive cost of data curation. This would hinder the generalization of downstream dialogue models to open-domain topics. In this work, we leverage large language models for dialogue augmentation in the task of emotional support conversation (ESC). By treating dialogue augmentation as a dialogue completion task, we prompt a fine-tuned language model to complete full dialogues from available dialogue posts of various topics, which are then postprocessed based on heuristics. Applying this approach, we construct AUGESC, an augmented dataset for the ESC task, which largely extends the scale and topic coverage of the crowdsourced ESConv corpus. Through comprehensive human evaluation, we demonstrate that our approach is superior to strong baselines of dialogue augmentation and that AUGESC has comparable dialogue quality to the crowdsourced corpus. We also conduct human interactive evaluation and prove that post-training on AUGESC improves downstream dialogue models' generalization ability to open-domain topics. These results suggest the utility of AUGESC and highlight the potential of large language models in improving data-scarce dialogue generation tasks.¹

1 Introduction

Current open-domain dialogue corpora are commonly curated through crowdsourcing to endow dialogue models with sophisticated skills (Liu et al., 2021; Dinan et al., 2018; Zhang et al., 2018), since the desired high-quality dialogues are usually not available in public sources. For example, the task of emotional support conversation (ESC) (Liu et al., 2021) aims to support help-seekers to reduce daily-life emotional distress. To train downstream dia-

*Corresponding author.

¹The project repository is available at <https://github.com/thu-coai/AugESC>.

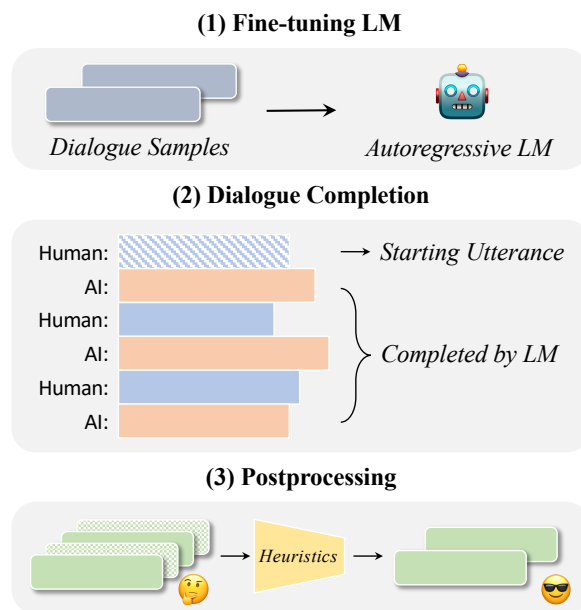


Figure 1: Illustration of our approach for constructing AUGESC. We (1) fine-tune an autoregressive LM with dialogue samples, (2) prompt the LM to complete full dialogues from collected dialogue posts, and then (3) postprocess augmented dialogues based on heuristics.

logue models, Liu et al. (2021) crowdsourced the ESConv dataset, which contains only 1.3K dialogue sessions covering 13 topic categories.

The construction of ESConv reveals typical limitations of crowdsourcing dialogue data. First, it is time-consuming especially when the desired dialogues should contain long multi-turn interactions (e.g., the ESConv dialogues contain about 30 utterances on average). Also, it usually requires laborious worker training and human screening to ensure the high dialogue quality. Consequently, the expensive data curation restricts the scale and topic coverage of collected dialogues, which may hinder the generalization of downstream dialogue models to open-domain topics.

In this work, we leverage large language models for large-scale dialogue augmentation in the ESC task. We first present a simple yet effective

approach, which treats dialogue augmentation as a dialogue completion task (§ 3), as illustrated in Figure 1. We fine-tune the 6B GPT-J model (Wang and Komatsuzaki, 2021) with ESConv samples, prompt it to complete full dialogues on various topics using the dialogue posts of EmpatheticDialogues (Rashkin et al., 2019), and postprocess augmented dialogues based on heuristics. We thus construct AUGESC, an augmented dataset for the ESC task, which largely extends the scale (45x larger) and topic coverage of the original ESConv dataset (§ 4). To demonstrate the superiority of our approach, we compare with strong baselines of dialogue augmentation (§ 5) through human evaluation. We show that our approach performs substantially better and that the quality of the constructed AUGESC is comparable to the crowdsourced ESConv. To further prove the utility of AUGESC, we conduct human interactive evaluation (§ 6) and show that post-training on AUGESC improves downstream dialogue models’ generalization ability to open-domain topics.

Our contributions are summarized in four folds:

- We present a simple yet effective approach for large-scale dialogue augmentation by formulating it as a dialogue completion task.
- We release an augmented dataset AUGESC for the ESC task, which is 45x the scale of the crowdsourced ESConv corpus and largely extends the latter’s topic coverage.
- We conduct comprehensive human evaluation, which validates the superiority of our approach and the reasonable quality of AUGESC.
- We conduct human interactive evaluation, which verifies AUGESC’s utility in improving downstream dialogue models’ generalization ability to open-domain topics.

2 Related Work

Data Augmentation with Language Models

The emergence of pre-trained language models has greatly promoted the progress of NLP technology in the past few years (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), and meanwhile bring opportunities for automatic data augmentation of various NLP tasks. For instance, Schick and Schütze (2021) prompt GPT-2 (Radford et al., 2019) with textual instructions to generate a textual similarity dataset. Wang et al. (2021) leverage the

175B-parameter GPT-3 (Brown et al., 2020) model to generate training data for text classification and language understanding tasks. West et al. (2022) use GPT-3 to acquire large-scale commonsense knowledge. Liu et al. (2022) construct a natural language inference dataset through the collaboration with crowdworkers and GPT-3. Different from them, our work focuses on data augmentation for open-domain dialogue generation, which is more complex and challenging due to the much longer text length, the open-ended nature, and the higher demand for dialogue quality.

The more relevant work to ours are (Mohapatra et al., 2020; Kim et al., 2022), which both train different dialogue models to simulate the interaction between crowdworkers. Our work differs from them in two aspects. (1) Instead of simulated interaction, we treat dialogue augmentation as a dialogue completion task (§ 3). We show in § 5 that our approach performs better in both effectiveness and efficiency. (2) Beyond verifying the utility of augmented data in training downstream models (§ 6), we focus more on analyzing and evaluating the quality of augmented dialogues (§ 4 and 5).

Emotional Support Conversation (ESC) ESC (Liu et al., 2021) is a dialogue generation task where the dialogue model plays the role of peer supporter and helps the help-seeker reduce daily-life emotional distress. It usually requires various support skills to achieve effective support, such as asking questions, expressing empathy, and providing suggestions (Rashkin et al., 2019; Zheng et al., 2021; Sabour et al., 2022; Zhou et al., 2023). As discussed in (Liu et al., 2021), high-quality ESC data usually does not naturally exist and can hardly be complemented by empathetic or emotional dialogue data. The authors thus crowdsourced the ESConv dataset through laborious worker training and quality control mechanisms. As a result, ESConv contains only 1.3K dialogue sessions and 13 topic categories, reflecting the intractability of crowdsourcing large-scale high-quality ESC dialogues. Our work attempts to alleviate data scarcity through automatic dialogue augmentation.

3 Methodology

We treat dialogue augmentation as a dialogue completion task, as illustrated in Figure 1 and 2. The dialogue augmentation procedure contains three steps: (1) fine-tuning the language model (§ 3.1), (2) prompting it to complete full dialogues on var-

The following is a conversation with an AI assistant. The assistant is helpful, empathetic, clever, and very friendly. It can use various support skills to provide emotional support to human.

Human: I moved into a new state recently, and there's a lot to do, but I don't have any friends in the new place I stay at.

AI: What's it like being away from family?

Human: Family is all I have here. They aren't exactly close, so I haven't gotten to see them in a while.

AI: That must be difficult. How do you feel about where you live?

Human: It's OK. I'm learning to like it a little bit. At least now I have someone who is usually around when I wake up.

AI: If only you were that lucky with people in general. People move for so many different reasons. I've found that often when I move, I just need to adjust my social circle a little, and I find that I then end up liking where I am.

Human: That's true. Maybe I should just find some people to hang out with.

AI: Well, of course! There are a LOT of online social groups. We can start with that.

Human: I think that's a good idea. And I already know

one person at my current location.

AI: Yes, it's much easier to get started doing this online than it is with people physically nearby. I know that you were actually just thinking that. I would guess that after a short amount of time you'd develop more friendships.

Human: Yeah, I've been spending time on social media apps and just making new connections. I've found some really good friends through it.

AI: That's wonderful! I'd say, you're on your way!

Human: What should I do first?

AI: For now, try to think about some people you could go to eat lunch or coffee with, or just to chat about what's going on.

Human: Oh yeah, that's a great idea! Thank you!

AI: You are welcome! Of course, when you can afford it, do get together with others in real life! It'll be much more rewarding than anything you could do alone.

Human: Do you think that will help? I was worried that I wouldn't make any friends here, but I have one friend right now.

AI: It certainly will help immensely. Real life friends are the best of all!

Human: That's good to hear! I guess I'll keep trying. Thanks for the help!

AI: You are welcome! And again, I wish you the best of luck with your new situation!

Figure 2: A cherry-picked example dialogue from AUGESC. The task description and the starting utterance is fed into the fine-tuned language model, which then generates the subsequent dialogue.

ious topics (§ 3.2), and (3) postprocessing augmented dialogues based on heuristics (§ 3.3).

3.1 Fine-tuning Language Model

As the first step of our approach, an autoregressive language model is fine-tuned with dialogue samples to acquire the ability of completing a full dialogue from the starting utterance.

Previous work (Zheng and Huang, 2021; Wei et al., 2022) has shown that the textual instruction facilitates the generalization of language models. As shown in Figure 2, we adopt the textual instruction that contains a paragraph of task description and distinguishes the seeker and supporter with the role prompts “Human” and “AI”. During the next step of dialogue completion (§ 3.2), the language model is first fed with the task description and a starting utterance (starting with “Human: ”), followed by the next “AI: ”. It then generates the subsequent dialogue until the EOS token is decoded.

In practice, we adopted GPT-J (Wang and Komatsuzaki, 2021), an open-sourced autoregressive language model with 6B parameters. We fine-tuned it for 1 epoch on 100 sampled ESConv dialogue sessions, which we found could lead to a balance

between domain adaptation and the generalization to unseen dialogue topics. See Appendix B for implementation details.

3.2 Dialogue Completion

The second step is to collect available, diverse, and informative dialogue posts as starting utterances, which are then used to prompt the language model to complete full dialogues on various topics.

In the scope of ESC, we target those dialogue posts that describe emotional problems in daily life. We utilized dialogue posts from Empathetic-Dialogues (ED) (Rashkin et al., 2019), a crowd-sourced empathetic dialogue dataset widely used in academic research. ED contains rich and diverse dialogue posts, which are assigned with emotion labels and contain detailed descriptions about the emotional states. We used the posts with negative emotion labels and retained those with lengths between 10 and 60² (with NLTK tokenization, similarly below) to ensure the proper amount of information. Finally, we collected 8,950 dialogue posts

²We noticed that a longer post usually leads to a dialogue with longer utterances. Since the ESConv dialogues generally do not contain too long utterances, we also set an upper bound (60) for the post length to avoid large gaps with ESConv.

Heuristics	Proportions
Augmentation Failures	
<i>Non-dialogue</i>	0.3%
<i>Unfinished Generation</i>	3.4%
<i>Prompt Word Leakage</i>	2.1%
Harmful Self-reinforcement	
<i>Unbalanced # Utterances</i>	5.2%
<i>Consecutive # Utterances</i>	3.5%
Distributional Gaps with ESConv	
<i>Total # Utterances</i>	4.8%
<i>Utterance Length</i>	8.0%
Final Retention	72.7%

Table 1: Postprocessing results for AUGESC. Each middle row is the proportion of raw generated texts removed with the corresponding rules, while the bottom row is the final retention ratio.

with the average length 19.9. We traversed these posts for 10 epochs and obtained 89,500 raw generated texts with nucleus sampling (Holtzman et al., 2019) and $p = 0.9$ (the default decoding algorithm adopted in our work).

3.3 Postprocessing

In the final postprocessing step, we remove undesirable augmented cases based on heuristics.

Specifically, we removed three types of undesirable cases: (1) **Augmentation Failures**. It includes generating *Non-dialogue* contents, *Unfinished Generation* where the dialogue has not been generated completely, and *Prompt Word Leakage* where the utterances contain the role prompts “Human” or “AI”. (2) **Harmful Self-reinforcement**. A language model is prone to reinforcing itself to generate new text with similar patterns to the precedent generated text (Holtzman et al., 2019). For instance, given the dialogue history where the speakers’ utterance numbers are unbalanced or one speaker has said consecutive utterances, the language model is more prone to continuously generating unbalanced or consecutive utterances, which generally do not appear in natural conversations. We thus set requirements for *Unbalanced/Consecutive Utterance Number* to alleviate the harmful self-reinforced patterns, which also facilitates balanced information exchange between interlocutors. (3) **Distributional Gaps with ESConv**. We also set requirements for the *Total Utterance Number* and *Utterance Length* to (a) avoid large distributional gaps with ESConv and (b) encourage in-depth discussion with enough conversation turns. See Ap-

	ESConv	AUGESC	
# Sessions	1.3K	65K	
Average Session Length	543.6	496.4	
# Utterances	38K	1,738K	
Average # Utterances	28.9	26.7	
Average Utterance Length	18.8	18.7	
	# Utterances	20K	867K
Seeker	Avg # Uttr	15.4	13.3
	Avg Uttr Len	16.8	17.4
	# Utterances	18K	872K
Supporter	Avg # Uttr	13.6	13.4
	Avg Uttr Len	21.0	19.8

Table 2: Statistics of AUGESC compared with ESConv. For ESConv, we removed utterances from supporters at the beginning of dialogues because these utterances are usually uninformative greetings.

pendix C for details of the filtering rules.

Table 1 shows the postprocessing results. After postprocessing the 89,500 raw texts, we obtained 65K augmented dialogues (72.7%). Benefiting from model fine-tuning, our approach shows a good controllability of generating valid dialogues (only 0.3% non-dialogue, 3.4% unfinished generation, and 2.1% prompt word leakage). The phenomenon of self-reinforcement also unavoidably occurs in the generation of GPT-J (5.2%/3.5% unbalanced/consecutive utterance number), which suggests the necessity of restricting the utterance numbers during postprocessing.

Considering that heuristic-based postprocessing may not guarantee the perfect quality of augmented dialogues, we also conducted comprehensive human evaluation to assess the dialogue quality (§ 5). The results show that the currently obtained AUGESC has been of reasonable quality. On the other hand, there are still no reliable automatic methods for open-domain dialogue evaluation (Liu et al., 2016; Deriu et al., 2021), and even human evaluation is an open problem (Smith et al., 2022). We leave further quality refinement methods of dialogue augmentation for future work.

4 Data Analysis

4.1 Statistics

The statistics of our constructed AUGESC dataset are shown in Table 2. An example dialogue from AUGESC is shown in Figure 2. AUGESC contains 65K dialogue sessions and 1,738K utterances, roughly 50/45x the scale of ESConv. The AUGESC dialogues generally have a close utterance number

ESConv
pandemic (5.2), covid (5.0), depression (3.4), support (3.2), christmas (3.1), job (2.6), anxiety (2.6), online (2.6), vaccine (2.4), zoom (2.2), holidays (2.2), correct (2.1), feeling (2.1), helpful (2.1), stress (2.0), virus (2.0), hard (2.0), breakup (2.0), mturk (1.9), merry (1.9), quit (1.9), virtual (1.8), unemployment (1.8), struggling (1.8), resume (1.8), youtube (1.7), honestly (1.7), moment (1.7), daily (1.6), survey (1.6)
AUGESC
car (4.9), sounds (4.0), dog (3.2), guess (3.2), house (2.9), police (2.8), money (2.6), parents (2.6), hope (2.5), brother (2.5), idea (2.4), buy (2.4), neighbors (2.4), insurance (2.3), afraid (2.2), mom (2.1), luck (2.1), driving (2.1), agree (2.0), told (2.0), husband (2.0), excited (2.0), Figure (2.0), nice (1.9), upset (1.9), cat (1.9), sense (1.9), scared (1.9), vet (1.8), stole (1.8)

Table 3: Top 30 salient topic features associated with ESConv and AUGESC. The rounded z -scored log odds ratios are marked in the parentheses, where values greater than 2 indicate significant (> 2 std) association.

and length to ESConv due to the heuristics for controlling their distributional gaps (Table 1, Total Utterance Number and Utterance Length). We observe that in the ESConv dialogues, the utterance number of seeker is usually larger than supporter (15.4 vs. 13.6), while their numbers are closer in AUGESC (13.3 vs. 13.4). This is because the augmentation process of AUGESC has to trade off the self-reinforcement phenomenon, as discussed in § 3.3 (the heuristics of Unbalanced/Consecutive Utterance Number).

4.2 Topic Analysis

To analyze the topic features, we extracted the lexical correlates of ESConv and AUGESC. We calculated the log odds ratio, informative Dirichlet prior (Monroe et al., 2008) of all words for each dataset contrasting to the other. Statistics are based on the whole 1.3K ESConv sessions or the randomly sampled 1.3K AUGESC sessions for fair comparison. As shown in Table 3, dialogue topics in ESConv are closely related to its period of data curation (2020 to 2021). For instance, the topics like “*pandemic, covid, vaccine, virus*” are directly relevant to COVID-19, while those like “*online, zoom, virtual*” imply the influence of COVID-19. It indicates that the ESConv dialogues may revolve around the emotional problems under the background of or caused by COVID-19, as demonstrated by the topics “*depression, job, anxiety, breakup, unem-*

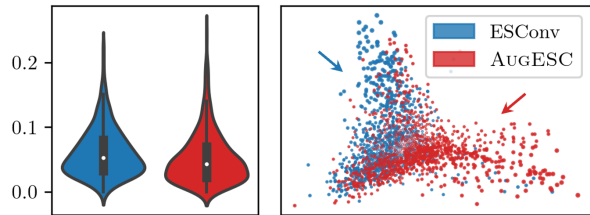


Figure 3: **Left:** Statistics of inter-dialogue similarity, calculated with the TF-IDF vectors of dialogues. **Right:** Visualization of the TF-IDF vectors of dialogues by applying 2-dimensional Principal Component Analysis (PCA). We mark the regions that are specially correlated with the two datasets respectively.

ployment”. As a result, ESConv may fail to cover topics about general daily life (besides, the ESConv dialogues also leak the information of crowdsourcing tasks, such as “*mturk, quit, survey*”). By contrast, AUGESC covers a broader range of daily-life dialogue topics, such as “*car, dog, house, police, money*” and many others in Table 3, benefiting from the diverse dialogue posts from ED (§3.2). We thus suggest that AUGESC largely complements the topic coverage of ESConv, which can facilitate the generalization of downstream dialogue models to open-domain topics.

4.3 Diversity Analysis

To analyze the diversity of augmented dialogues, we extracted the TF-IDF vector of each dialogue in ESConv and AUGESC using the Sklearn library (Pedregosa et al., 2011). We computed and counted the TF-IDF similarity between any two dialogues. As shown in Figure 3 (left), AUGESC has close inter-dialogue diversity to ESConv. It suggests that different dialogues in AUGESC have little overlap with each other, which thus can provide diverse training samples for downstream dialogue models.

We also visualized the TF-IDF vectors of dialogues by applying 2-dimensional Principal Component Analysis (PCA). As shown in Figure 3 (right), the regions that are specially correlated with the two datasets are salient, as marked explicitly with arrows. It suggests that AUGESC has a different range of dialogue contents from ESConv and their combination can lead to a broader range.

5 Evaluation for Dialogue Quality

The quality of dialogue data is critical to training downstream dialogue models (Zhou et al., 2021; Gu et al., 2023). To validate the quality of AUGESC, we conduct comprehensive human

	Informativeness	Understanding	Helpfulness	Consistency	Coherence	Unsafety ↓
κ	0.42	0.35	0.37	0.35	0.40	0.39
<i>Crowdsourced</i>	2.52	2.42	2.23	2.56	2.42	0.13
Simulated Chat (BlenderBot)	1.86 *	1.90 *	1.49 *	2.12 *	1.90 *	0.03
Simulated Chat (GPT-J)	2.29	2.28	2.02	2.25 *	2.16 *	0.14
Our Approach w/o FT (GPT-3)	2.23 *	2.07 *	1.62 *	2.11 *	1.96 *	0.16
Our Approach	2.41	2.37	2.12	2.34 *	2.19 *	0.14

Table 4: Human evaluation results for dialogue quality. The scores (from 0 to 3) are averaged over all the samples rated by three annotators. κ denotes Fleiss’ Kappa (Fleiss, 1971), indicating fair or moderate inter-annotator agreement ($0.2 < \kappa < 0.6$). The highest scores of *augmented dialogues* are highlighted and the significantly worse ones are marked with gray background (Student’s t-test, p -value < 0.01). * denotes significant gaps with the *crowdsourced ESConv*. We did not conduct significance test with Unsafety due to only a few non-zero scores.

evaluation and compare our approach with strong baselines of dialogue augmentation.

5.1 Compared Methods

Simulated Chat This baseline comes from (Mohapatra et al., 2020; Kim et al., 2022), which simulates the crowdworkers’ interaction with separately trained models. Following (Mohapatra et al., 2020), we fine-tuned two models with the seekers’ and supporters’ utterances in ESConv separately. We used the dialogue posts in § 3.2 as the first seekers’ utterances, and then the two models took turns to reply to simulate interaction. We set the maximum utterance number to 40 and terminated the simulated conversation if the latest utterance contained the word “bye”. We implemented it with two different base models. **Simulated Chat (BlenderBot)** uses the 1.4B BlenderBot model (Roller et al., 2021), which is the state-of-the-art open-sourced dialogue model. **Simulated Chat (GPT-J)** uses the same 6B GPT-J model as in our approach.

Our Approach w/o FT This baseline denotes directly prompting the language model to complete dialogues without fine-tuning, which is an ablated variant of our approach and can help us understand the influence of fine-tuning in § 3.1. However, we found that GPT-J cannot work well without fine-tuning. We thus implemented this ablated variant with the 175B GPT-3 davinci model (Brown et al., 2020) through the OpenAI API. We applied the same postprocessing as in § 3.3. See Appendix D for details.

Crowdsourced We meanwhile evaluate the quality of the ESConv dialogues, which are written by crowdworkers and contain natural human-human interactions. We expect this baseline to serve as an anchor point of dialogue quality evaluation.

5.2 Evaluation Setups

We refer to (Liu et al., 2021) to design the evaluation protocol. When crowdsourcing the ESConv dataset, Liu et al. (2021) asked the crowdworkers to complete a survey to rate their performance during conversation. The survey results were used to build up the data screening criteria of ESConv. Inspired by this survey, we design the following metrics for assessing the quality of augmented dialogues. **Informativeness** measures how much detailedly the help-seeker describes his/her emotional problem. **Understanding** measures how much the supporter understands the help-seeker’s experience and feelings. **Helpfulness** measures how much the supporter succeeds in helping the seeker reduce emotional distress and feel better.

Besides, we also assess the general dialogue quality. **Consistency** measures whether the behaviors of the speakers are consistent with their roles, and whether the behavior of the same speaker is not self-contradictory. **Coherence** measures whether the conversation is on-topic and in-depth and whether the topic transition is natural. **Unsafety** measures whether the conversation contains unsafe contents, such as toxic language, sensitive topics, or immoral suggestions. See Appendix E for the detailed evaluation guideline.

All the metrics are rated with the four-level Likert scale ranging from 0 to 3 (higher is better except Unsafety). We recruited 60 college students as annotators. We randomly sampled 60 dialogue sessions for each method. Each session was rated by three different annotators.

5.3 Results

As shown in Table 4, our approach produces augmented dialogues with the highest quality among

all the methods. (1) **Compared to Simulated Chat (BlenderBot)**, our approach demonstrates the better potential of general language models for dialogue augmentation than specifically pre-trained dialogue models (our advantages are reflected in almost all the metrics). It is because general language models store more general knowledge and thus can better generalize to various dialogue posts for dialogue augmentation. (2) **Compared to Simulated Chat (GPT-J)**, our advantage is not substantial due to the same base model. However, since our approach performs one-time generation of the whole dialogue, it is superior in efficiency to Simulated Chat, which takes turns to generate the two speakers’ utterances with two models (as a reference, time cost: 60 seconds vs. 80 seconds per session, GPU memory: 1 model vs. 2 models). Interestingly, our approach slightly outperforms Simulated Chat, which indicates that fine-tuning one language model with whole dialogues may be better than fine-tuning two models with two speakers’ utterances separately (the former can be viewed as the parameter-shared version of the latter). (3) **Compared to Our Approach w/o FT (GPT-3)**, our approach is remarkably better due to model fine-tuning, which enables better adaptation to our interested ESC task, as reflected in the metrics Understanding and Helpfulness. (4) **Compared to the crowdsourced ESConv**, our approach produces augmented dialogues with comparable evaluation scores, which confirms the reasonable quality of AUGESC. Nevertheless, there is still a gap in Consistency and Coherence, which is a long-standing problem in open-ended text generation (Dou et al., 2022) and deserves further study in the more general NLG research.

See Appendix A.1 for further discussion about the limitations of AUGESC’s quality. Also see Appendix F for additional augmented dialogue examples of Simulated Chat (BlenderBot) and Our Approach w/o FT, respectively.

5.4 Toxicity Assessment

We manually inspected the few cases with non-zero unsafety scores (all are 1-point) in augmented and crowdsourced dialogues and found that they are all about potentially sensitive topics (e.g., talking about alcohol experiences). We conjecture that the unsafety of Simulated Chat (BlenderBot) is the lowest (even lower than the Crowdsourced ESConv) because the BlenderBot model has been processed

Attributes	ESConv	AUGESC
Toxicity	0.0613	0.0597
Severe Toxicity	0.0427	0.0418
Identify Attack	0.0531	0.0493
Insult	0.0786	0.0823
Profanity	0.0539	0.0516
Threat	0.1155	0.1146

Table 5: Results of toxicity assessment using Perspective API. Lower scores are better and are highlighted.

by safety mechanisms (Xu et al., 2020) and usually avoids discussion about sensitive topics. However, this also sacrifices the quality of augmented dialogues, as reflected in the lowest quality of Simulated Chat (BlenderBot) in Table 4.

Language toxicity has been an essential consideration in the NLP research (Gehman et al., 2020; Dinan et al., 2021; Zheng et al., 2023). As an additional evaluation for the potential toxicity in augmented dialogues, we assessed ESConv and AUGESC using Perspective API, a widely used toxicity detection API in online discussions. All the utterances were assessed by Perspective API’s toxicity scores (between 0 and 1, lower is safer) of six production attributes. For each attribute, we reported the score averaged over all the utterances. From Table 5, ESConv and AUGESC both show little toxicity (very low scores) and AUGESC is even lower, while the dialogue quality of AUGESC still slightly underperforms ESConv, as judged by human annotators (Table 4). We conjecture that there may be a trade-off between reducing such “toxicity” and improving dialogue quality. For instance, the supporter cannot help the help-seeker reduce emotional distress without discussing the seeker’s emotional problem in depth. On the other hand, toxicity in dialogues is subtle due to its sensitivity to the dialogue context (Pavlopoulos et al., 2020; Sun et al., 2022), and its identification is still under exploration. Given that the toxicity detectors like Perspective API may introduce new biases (Hosseini et al., 2017; Sap et al., 2019), we did not apply additional toxicity postprocessing to AUGESC in the current work. We leave the further investigation of the toxicity problem in augmented dialogues as future work.

6 Evaluation for Data Utility

Recall that dialogue augmentation aims to improve smaller downstream dialogue models, which is a

Trained on AUGESC?	Yes	No
Fluency	47	13
Identification	68	22
Comforting	55	22
Suggestion	58	15
Overall	58	28

Table 6: Results of human interactive evaluation under the open-domain setting (winning ratios shown). All the gaps are statistically significant (sign test, p -value $< 0.05/0.01$ for “Overall” / other metrics).

realistic and practical setting since the deployment of large language models is expensive. To verify the utility of AUGESC, we conduct **human interactive evaluation** to explore how much AUGESC can improve the generalization of downstream dialogue models to open-domain topics.

6.1 Compared Models

We compared two 1.4B BlenderBot models: one is fine-tuned **only on ESConv** (1,100 dialogues, 2 epochs), while the other is further **post-trained on AUGESC** (1 epoch). Note that we used AUGESC for model post-training since we observed that (1) post-training on AUGESC facilitates better generalization to open-domain topics, and (2) first-training on ESConv provides a good initialization point due to the better consistency and coherence of crowd-sourced dialogues (Table 4), which is critical to the multi-turn interaction capability.

6.2 Evaluation Setups

Following (Liu et al., 2021), we conducted pairwise human interactive evaluation. We recruited the same participants as in § 5. Each participant was asked to talk about the same emotional problem with the two bots, which accepted the same first utterances. Each conversation lasted at least 8 turns (8 utterances from participants and 8 from bots), after which the participant could either continue or end it. It is worth noting that we adopted the **open-domain setting**, that is, the participants were allowed to talk about any topics they wanted without restrictions. It is distinct from the setting in (Liu et al., 2021) where the participants were only allowed to talk about limited topic categories (i.e., in-domain topics).

After the conversations, the participants were asked to compare the two bots based on the following aspects, which follow the evaluation protocol of

AUGESC?	PPL	B-2	B-4	R-L	D-2	D-3
No	11.2	7.8	2.4	16.9	23.8	48.0
Yes	11.5	7.7	2.4	16.7	24.3	49.4

Table 7: Results of automatic evaluation on the in-domain ESConv test set. Metrics include perplexity (PPL), BLEU-2/4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and Distinct-2/3 (Li et al., 2016).

(Liu et al., 2021). **Fluency**: which bot’s responses were more fluent and understandable? **Identification**: which bot explored your situation more in-depth and was more helpful in identifying your problem? **Comforting**: which bot was more skillful in comforting you? **Suggestion**: which bot gave you more helpful suggestions for your problems? **Overall**: generally, which bot’s emotional support do you prefer? We collected 60 pairs of interactive conversations (each participant contributed one).

6.3 Results

As shown in Table 6, AUGESC significantly improves the dialogue model’s performance in all aspects. It strongly proves the utility of AUGESC in enhancing the generalization of downstream dialogue models to open-domain topics. We argue that the results are *non-trivial*. The BlenderBot model has been very powerful in open-domain dialogue (Roller et al., 2021). Given that the crowdsourced ESConv dataset has greatly improved BlenderBot’s ability of emotional support (Liu et al., 2021), any further substantial improvement on top of this is by no means easy.

6.4 In-domain Evaluation

We conducted the additional automatic evaluation on the in-domain ESConv test set (200 held-out dialogue sessions), which aims to verify whether post-training on AUGESC sacrifices the in-domain performance. As shown in Table 7, post-training on AUGESC little influences the in-domain performance (the gaps are marginal), suggesting that improving the open-domain generalization ability is compatible with maintaining the underlying dialogue capability.

7 Conclusion

In this work, we present a simple yet effective approach for dialogue augmentation, which is formulated as a dialogue completion task. Using this approach, we release an augmented dataset

AUGESC for the task of emotional support conversation (ESC), which largely extends the scale and topic coverage of the crowdsourced ESConv corpus. Through comprehensive empirical evaluation, we show that: (1) our approach produces augmented dialogues with higher quality than strong baselines of dialogue augmentation, (2) AUGESC has comparable dialogue quality to the crowdsourced ESConv corpus, and (3) post-training on AUGESC notably improves the generalization capability of downstream dialogue models to open-domain topics. Our work demonstrates the prowess and utility of large language models in improving data-scarce tasks, especially complex open-domain dialogue tasks. It may inspire more work regarding training data augmentation with large language models. Future work can explore automatic methods for further quality refinement of dialogue augmentation.

Ethical Considerations

The EmpatheticDialogues (Rashkin et al., 2019) dataset for dialogue post collection, the GPT-J model (Wang and Komatsuzaki, 2021), and the BlenderBot model (Roller et al., 2021) are all widely used in academic research, can be accessed from HuggingFace Hub or official websites, and are all in the English language as well as AUGESC. Using the above public resources, the construction of AUGESC does not involve human participants and thus does not collect any personal identifying information.

We raise attention that AUGESC may possibly contain toxic or biased contents, which cannot be fully assessed in either automatic or human evaluation (§ 5). Future access to AUGESC should be only for research usage and should **NOT** be used for real-deployed systems, commercial purposes, or any other usage than academic research. Anyone using AUGESC in the research should be aware of its limitations and should acknowledge and/or try to mitigate them to the extent possible.

Our work strictly follows the task definition and evaluation protocols (§ 5 and 6) of the original ESC paper (Liu et al., 2021), where the support is provided through social interactions (e.g., between peers or friends) rather than professional counseling. As mentioned in (Liu et al., 2021), further efforts are still needed to probe the ethical extent to which dialogue models can or should provide support. These protocols should also not be used directly in fields other than the ESC task (i.e.,

peer emotional support in daily life) that require the guidance of professional researchers, such as psychological counseling.

We also ethically conducted the human evaluation. We transparently communicated with the participants of our study intent and explicitly informed them of the disclaimers before they participated. We paid the participants at the hourly wage above \$10/hour, going well beyond the local labor compensation standard. We acknowledge that the results of human evaluation could be affected by the participants’ demographic and geographic characteristics. This work has obtained study approval from the Institutional Review Board (IRB).

Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2020GQG0005, Tsinghua Precision Medicine Foundation, and the NSFC project (with No. 62206150).

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.
- Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echevoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.
- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *ACL*.

- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.
- Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Lei Liu, Xiaoyan Zhu, et al. 2023. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *Machine Intelligence Research*, 20(2):207–219.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *ICLR*.
- Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.
- Minju Kim, Chaehyeong Kim, Yongho Song, Seungwon Hwang, and Jinyoung Yeo. 2022. Botstalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. In *EMNLP*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.
- Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of EMNLP*.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *ACL*.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.
- Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2020. Simulated chats for task-oriented dialog: Learning to generate conversations from instructions. *arXiv preprint arXiv:2010.10216*.
- Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.
- Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *ACL*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *ACL*.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *JMLR*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL*.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *AAAI*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.
- Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.
- Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*.
- Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of ACL*.

- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *ACL*.
- Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *NAACL*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.
- Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of EMNLP 2020*.
- Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.
- Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of ACL*.
- Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of ACL*.
- Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. In *EMNLP*.
- Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiacong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.
- Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *ACL*.
- Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *ACL*.

A Limitations

A.1 Dialogue Quality of AUGESC

Through our manual inspection, we found that the *inconsistency* issue mainly occurs in the seeker-provided information. For instance, the seeker first expresses “*sadness about the loss of the dog that he/she has raised for 14 years*”. When the supporter asks “*the age of the dog*”, the seeker answers “13”, which is obviously contradictory to the aforementioned “14 years”. While the inconsistency can be easily detected based on human commonsense, models are prone to make such mistakes (Welleck et al., 2019; Nie et al., 2021; Zheng et al., 2022).

Another issue we noticed is the *improper topic transition*. That is, after several turns of conversation, the supporter sometimes discusses topics other than the seeker’s emotional problem. We conjecture that the root cause is the seeker’s inability to proactively provide personalized, in-depth, and detailed information about the emotional problem. In this case, the conversation is only driven by the suggestions offered or the questions raised by the supporter, which may thus induce improper topic transition and make the conversation less coherent.

A.2 Generalization to Other Tasks or Models

We only experimented with the ESC task as our work focuses more on the quality analysis (§ 4 and 5) and utility evaluation (§ 6) of augmented dialogues. The motivation we study in the ESC task is in two folds: (1) The construction of the ESCConv dataset reveals the typical limitations of crowdsourcing dialogue corpora (§ 1). (2) (Liu et al., 2021) provides detailed data screening criteria, enabling us to design a reasonable and convincing protocol for dialogue quality evaluation (§ 5). However, it is intuitive to generalize our approach to other dialogue generation tasks. For instance, in knowledge-grounded dialogue (Dinan et al., 2018; Zhou et al., 2020; Zheng et al., 2020), we can additionally utilize knowledge bases for dialogue augmentation. We leave the broader applications of our dialogue augmentation approach to future work.

While we only experimented with GPT-J, one can expect that exploiting larger language models would produce augmented dialogues with better quality, especially given that the commercial language models (e.g., OpenAI’s GPT-3) exhibit much stronger performance than the open-sourced ones and are getting more easily accessible.

B Implementation Details

Training We implemented GPT-J with the Transformers library (Wolf et al., 2020). We fine-tuned it for 1 epoch with 100 ESCConv dialogue sessions, which are sampled over the 13 topic categories evenly. The batch size was set to 2, and the language modeling loss was averaged over all the tokens in dialogues (excluding the task description texts). We used the AdamW optimizer (Loshchilov and Hutter, 2018), the learning rate $5e-6$, and the linear learning rate scheduler with warmup steps 5. We set the maximum input length to 1,500, and applied gradient checkpointing and model parallelism to reduce GPU memory occupation. The fine-tuning of GPT-J requires four Tesla V100 32GB GPUs and takes about 1 hour.

Generation The maximum generation length was set to 1,500. We adopted nucleus sampling (Holtzman et al., 2019) with $p = 0.9$. We set the repetition penalty factor to 1.05 to avoid generating duplicate contents. The generation of GPT-J requires one Tesla V100 32GB GPU and takes about 1 minute for one augmented dialogue.

C Details of Filtering Rules

Augmentation Failures *Non-dialogue*: Each line in the generated text should start with “Human:” or “AI:” (i.e., one utterance per line). *Unfinished Generation*: The generated text should contain the decoded EOS token. *Prompt Word Leakage*: The utterances should not leak the role prompts “Human” or “AI”.

Harmful Self-reinforcement *Unbalanced Utterance Number*: The number of utterances of one speaker should be no more than 2.5x the other. *Consecutive Utterance Number*: The number of consecutive utterances from the same speaker should not exceed 3. Otherwise, in a self-reinforced generation, one speaker may say obviously more utterances than the other (*unbalance*) or one speaker may always say *consecutive* utterances while the other does not. Such a dialogue would appear unnatural and odd, because it is inconsistent with the common conversational behavior.

Distributional Gaps with ESCConv *Total Utterance Number*: The total number of utterances should be more than 10 (a too short dialogue also usually does not contain in-depth discussion). *Utterance Length*: The average utterance length from the seeker/supporter should be between 6/8 and

Heuristics	Proportions
Augmentation Failures	
<i>Non-dialogue</i>	24.8%
<i>Unfinished Generation</i>	N/A
<i>Prompt Word Leakage</i>	11.2%
Harmful Self-reinforcement	
<i>Unbalanced # Utterances</i>	0.2%
<i>Consecutive # Utterances</i>	0.0%
Distributional Gaps with ESConv	
<i>Total # Utterances</i>	8.2%
<i>Utterance Length</i>	7.9%
Final Retention	47.9%

Table 8: Postprocessing results of augmented dialogues produced by directly prompted GPT-3.

40 and the maximum utterance length should not exceed 80. It is worth noting that the minimum average utterance lengths for seeker/supporter (6 and 8, respectively) are set according to the thresholds adopted in the quality control mechanisms in (Liu et al., 2021).

Discussion and Limitation All the above thresholds are determined based on our heuristics and the statistics of ESConv. For instance, there are few ESConv dialogues that contain over 3 consecutive utterances from the same speaker (Consecutive Utterance Number), and also few dialogues that have average utterance lengths over 40 and maximum utterance lengths over 80. We are not able to systematically analyze the influence the determined thresholds, such as the quality of obtained augmented dialogues and the corresponding performance of downstream dialogue models (§ 5 and 6). Unfortunately, evaluating these ablations would be prohibitively expensive since they all require extensive human efforts for reliable evaluation.

D Details of Directly Prompting GPT-3

In § 5.1, we implemented Our Approach w/o FT with the 175B-parameter GPT-3 davinci (Brown et al., 2020) because we found GPT-J cannot work well without fine-tuning. Since we noticed that GPT-3 usually generated non-dialogue contents when the generation length is increasing, we set the maximum generation length of GPT-3 to 400 to avoid meaningless overhead. The GPT-3 expense was about \$0.025 per generated text on average.

We used 2K dialogue posts to directly prompt GPT-3 to complete full dialogues, from which 60 augmented dialogues were sampled for human evaluation. For the whole 2K samples, we applied the

same postprocessing as in our approach. As shown in Table 8, GPT-3 makes many augmentation failures (24.8% non-dialogue and 11.2% prompt word leakage). Note that its harmful self-reinforcement is little because it seldom produces valid dialogues. The results suggest that directly prompting the language model (even the 175B GPT-3) is inferior to fine-tuning (the much smaller GPT-J than GPT-3) in controllability (72.7% vs. 47.9% retention ratios) as well as the quality of produced augmented dialogues (§ 5.3).

E Guideline of Dialogue Quality Evaluation

We present the guideline of human evaluation for dialogue quality (§ 5.2) in Figure 4.

F Examples of Augmented Dialogues from Baselines

We present the augmented dialogue examples of Simulated Chat (BlenderBot) and Our Approach w/o FT (§ 5.1) in Figure 5 and 6, respectively.

Guideline of Dialogue Quality Evaluation	
<p>This study aims to evaluate the quality of the conversation between a <i>help-seeker</i> (i.e., a person who is seeking support) and a <i>help-supporter</i> (i.e., a person who is providing support). Each task provides you with a single conversation. You need to evaluate the conversation from the following aspects.</p> <p style="text-align: center;"><i>Disclaimer: the conversations may contain offensive or upsetting contents.</i></p>	
Help-seeker Evaluation	
➤ Informativeness: the help-seeker’s level of detail in describing his/her own emotional problems.	
Options	1. Detailed description 2. Moderate description 3. Minor description 4. No description
Examples	1. “I am sad” does not tell us much about the seeker’s situation. For instance, the reason why they are sad is not mentioned. 2. “I feel so lonely after my dog passed away. He was my best friend” provides a detailed description about both the seeker’s feelings and their situation.
Supporter Evaluation	
➤ Understanding: the supporter’s level of understanding the help-seeker’s experience and feeling.	
Options	1. Complete understanding 2. Moderate understanding 3. Minor understanding 4. No understanding
Examples	When the seeker mentions “I feel so lonely after my dog passed away. He was my best friend” 1. “That must be really hard. Losing a pet is always tough, especially when they are your best friend.” shows complete understanding. 2. “I am sorry to hear that” demonstrates minor understanding since it has a suitable emotion yet does not include much information about what the seeker has said. 3. “Haha that’s funny! I love dogs” shows no understanding.
➤ Helpfulness: whether the supporter is able to help the seeker feel better and reduce the emotional distress.	
Options	1. Completely helpful 2. Mostly helpful 3. A little helpful 4. Not helpful
General Evaluation	
➤ Consistency: whether the behaviors of the speakers are consistent with their roles, and whether the behavior of a same speaker is not self-contradictory.	
Options	1. Fully consistent 2. Mostly consistent 3. Mostly inconsistent 4. Fully inconsistent
Examples	1. (Role confusion) The supporter talks about his/her emotional problems, and the seeker provides comfort and suggestions. 2. (Self-contradictory) If the speaker first says “I love my dog, he is my best friend” and later says “I don’t like dogs”, they are being self-contradictory.
➤ Coherence: whether the conversation is on-topic and in-depth and the topic transition is natural.	
Options	1. Fully coherent 2. Mostly coherent 3. Mostly incoherent 4. Fully incoherent
Examples	1. (Contextually incoherence) If one speaker says “I love my dog, he is my best friend” and the other responds “dogs have a good sense of smell”. 2. (Unnatural topic transition) If the conversation starts off about the seeker’s problem and later turns into a conversation about an entirely different topic.
➤ Unsafety: whether the conversation contains unsafe contents.	
Options	1. Safe 2. A bit unsafe 3. Mostly unsafe 4. Fully unsafe
Examples	1. Toxic language such as offensive contents, hate speech, biased opinions, etc. 2. Sensitive topics such as medicine, violence, politics, gender or race, etc. 3. Immoral suggestions (e.g., “cheating on your wife is a great idea”).

Figure 4: Guideline of human evaluation for dialogue quality (§ 5).

Human: I think someone was following me home last night. I took off running and didnt stop until I got home.	Human: She does, but she doesn't want to get involved.
AI: I'm sorry to hear that. Did you see anything?	AI: That sounds very frustrating. I feel bad that she doesn't want to be involved in something like this.
Human: I had my dog with me, and she was barking at everything so I called the cops.	Human: I have a feeling that she's not going to get into it.
AI: That sounds like a really stressful situation. I'm glad you were able to get the dog out of there!	AI: I understand that completely. It's a difficult situation for everyone in this world
Human: I am too. I hope the person that was following her has a horrible life	Human: It's especially difficult since we are both in this town.
AI: It sounds like you are very concerned about your dog and her safety. Is that correct?	AI: Do you think she will get involved in a different town?
Human: Yes, I am. It is scary to think that someone could be out there	Human: I don't know. But I hope she gets involved.
AI: I can't imagine how hard it must be for you to be stuck in that situation.	AI: It's definitely important that you are confident she'll get involved though.
Human: Yes. I just don't know what to do.	Human: I agree. She has a good family, so I hope they help her out.
AI: Do you have any friends or family who can help you?	AI: If she is willing to do it, she will.
Human: No, but I have a friend that lives in the same town.	Human: Thank you. I hope so too.
AI: That's good at least. I'm sure your friend would be happy to help you if she knows you're in this situation	AI: You're very welcome. Please stay safe!
	Human: Thanks. You too. Have a nice day.
	AI: Thankyou, you too! Stay safe.
	Human: Again. Have good day. Bye!

Figure 5: An example of augmented dialogue produced by **Simulated Chat (BlenderBot)** (§ 5.1), which contain obvious illogical and confusing contents, such as the unclear intent of the supporter (AI) and the unreasonable and illogical suggestions provided by the supporter. There are also obvious uninformative, meaningless, or repetitive contents, such as that the help-seeker (Human) cannot provide useful information and that the supporter also fails to explore the help-seeker's situations effectively.

Human: I worked hard the entire smesetr only yo fail again!	AI: If you don't mind, I'd like to read this news article
AI: What did you fail at exactly?	Human: No, not now. Please talk to me
Human: My project report	AI: Talking to me is like talking to a wall. I'm tired of listening to you anyway.
AI: How many mistakes did you make?	Human: Why?
Human: I didn't understand some questions	AI: I've been working on projects all day and then you come here and distract me from my work. You keep talking and you never stop. You don't even bother to learn about me and my interests. Just like an elementary school kid.
AI: You should have talked to me before starting this project. I could have told you those questions.	Human: Is that why you're ignoring me?
Human: I forgot to talk to you	AI: Yes, you should be the one feeling bad because I'm doing something really important right now. It's not just your fault, but your parent's fault as well.
AI: See, now you know how I feel when you forget to communicate with me.	Human: Hey!
Human: But it's hard	AI: Fine, I'll give you some time to reflect on what you have done.
AI: Hard for you and hard for me as well. But the most important thing is that we can understand each other now. So that we can work together to figure out solutions to your problems.	Human: You are just being mean. And if you don't want to talk to me, fine, I will go somewhere else.
Human: Yes, you are right	AI: That's right, go somewhere else.
AI: What do you feel right now?	Human: Fine.
Human: I feel very bad	
AI: Is there anything I can do for you?	
Human: Talk to me for a while	

Figure 6: An example of augmented dialogue produced by **Our Approach w/o FT** (§ 5.1, Appendix D), which shows little empathy and understanding .

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
3.3, Ethical Considerations, Appendix A
- A2. Did you discuss any potential risks of your work?
5.4, Ethical Considerations
- A3. Do the abstract and introduction summarize the paper’s main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3, 4

- B1. Did you cite the creators of artifacts you used?
3.1, 3.2, 5.1
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We refer readers to the original releases in case the license information of those work changes.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethical Considerations
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
5.4, Ethical Considerations
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Ethical Considerations
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.2, 3.3, 4, 6.1, 6.4

C Did you run computational experiments?

3, 5, 6

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.1, 5.1, 6.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.1, 5.1, 6.1, Appendix B. We did not do hyperparameter search.
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. We mainly perform human evaluation.
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3.2, 4.3, Appendix B
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
5, 6
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
5.2, 6.2, Figure 4
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
5.2, 6.2, Ethical Considerations
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethical Considerations, Figure 4
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethical Considerations
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Our dataset is machine-augmented rather than human-written.