Learning from Implicit User Feedback, Demographic Information and User Emotions in Task-Oriented Document-Grounded Dialogues

Anonymous ACL submission

Abstract

Trustworthiness, interaction quality and empathy have a great influence on whether users accept a dialogue system. To address this, recent works on open-domain dialogues suggest to learn from implicit user feedback or to consider demographic information and user emotions in response generation to improve generation accuracy and user engagement. However, for task-oriented and document-grounded dialogue systems, task completion and factual consistency of the generated responses are almost more important. The impact of such data on these quality criteria is not yet known. To address this gap, we (1) introduce FEDI, the first English task-oriented document-grounded dialogue dataset annotated with implicit user feedback, demographic information and user emotions, and (2) investigate the impact of including such data on task completion, and the factual consistency of responses generated by Flan-T5, GPT-2, and Llama 2. Our results show a particularly positive impact on task completion and factual consistency, and that responses generated by models trained with implicit user feedback are preferred by human users.[1]

1 Introduction

Trustworthiness, interaction quality and empathy have a great influence on whether users accept a dialogue system (Pelau et al., 2021). In this respect, the ability to recover from generation errors, using culturally familiar communication styles and being empathetic with users are key characteristics (Minjin Rheu and Huh-Yoo, 2021; Chaves and Gerosa, 2021). For example, in the second utterance of Figure 1, the system misinterprets Claudia's question resulting in an incorrect response that affects her emotional state and she now asks the system for clarification. After the system has generated a more suitable response, she is satisfied, her emotional state changes again and she asks the system another question. To address this in response generation, learning from implicit user feedback (Hancock et al., 2019; Veron et al., 2021; Xu et al., 2023b), such as a correction or question in response to a generation error in the preceding system utterance, considering demographic information (Lee et al., 2022; Zhang et al., 2018), such as age, occupation or language style, and user emotions (Rashkin et al., 2019; Hsu et al., 2018; Hwang et al., 2023) are promising approaches for improving the generation accuracy and user engagement in open-domain dialogue systems. However, for task-oriented and document-grounded dialogue systems, task completion and factual consistency of the generated responses, are almost more important (Budzianowski et al., 2018; Nekvinda and Dušek, 2021; Honovich et al., 2021). Although some work is available for learning from implicit

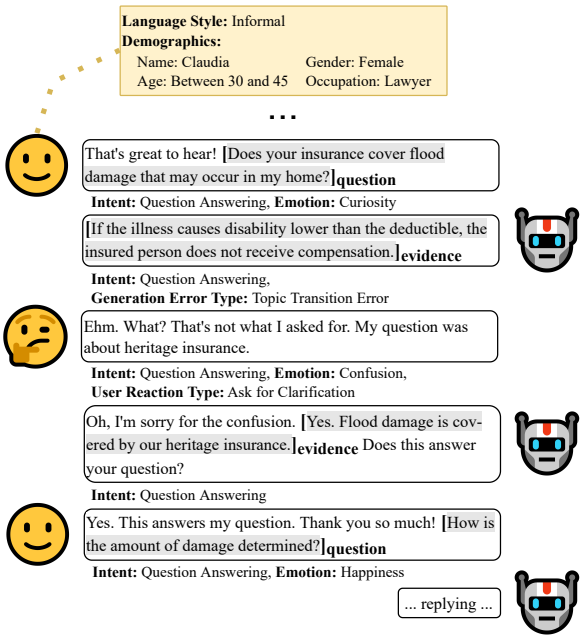

Figure 1: A feedback dialogue from FEDI. User emotion and implicit user feedback annotations (generation error and user reaction types) are beneath the utterances.

[1] Code and data are available in [placeholder].

Dataset	Source	Type	Demographic Information	User Emotions	Implicit User Feedback	#Dialogues	Avg. Num. of Turns	Avg. Utt. Length	Lexical Diversity
EmoWOZ (Feng et al., 2022)	Crowdsourced	Task-Oriented		✓		12k	9.5	8.2	55.7
FITS (Xu et al., 2023b)		Document-Grounded			✓	22k	7.1	15.0	52.8
Blenderbot 3x (Xu et al., 2023a)		Open-Domain			✓	261k	11.3	14.2	47.3
SaferDialogues (Ung et al., 2022)					✓	8k	2.5	14.8	53.3
EmotionLines (Hsu et al., 2018)				✓		1k	7.3	7.8	68.5
EmpatheticDialogues (Rashkin et al., 2019)				✓		25k	4.3	13.7	64.2
SODA (Kim et al., 2023)	LLM-Generated	Open-Domain		✓		1.5M	7.6	16.1	68.0
PersonaChatGen (Lee et al., 2022)			✓			1.6k	16.0	9.5	56.7
FEDI	**LLM-Generated**	**Task-Oriented Document-Grounded**	✓	✓	✓	**8.8k**	**7.6**	**16.8**	**62.1**

Table 1: Comparison of FEDI to other datasets that provide related annotations. FEDI is comparable with other LLM-generated datasets in terms of avg. turn and utterance length, and has a higher lexical diversity than many of the crowdsourced datasets[2].

user feedback in such systems (Wang et al., 2019; Veron et al., 2021; Mazumder et al., 2020), the impact of such data on these quality criteria is not yet known.

In this work, we address this gap by (1) introducing FEDI, the first English task-oriented document-grounded dialogue dataset annotated with implicit user **F**eedback, **E**motions and **D**emographic **I**nformation, and (2) investigating the impact of including such data on task completion and factual consistency of the generated responses using three state-of-the-art language generation models, i.e., Flan-T5 (Chung et al., 2022), GPT-2 (Radford et al., 2019) and Llama 2 (Touvron et al., 2023b).

We use GPT-3.5-Turbo[3] to generate and annotate the training and validation data of FEDI, and recruit humans to assess its quality and to collect a separate set of test dialogues. In summary, we provide these contributions:

1. FEDI, the first task-oriented document-grounded dialogue dataset for learning from implicit user feedback, demographic information and user emotions.

2. New experimental insights showing that including such data has a positive impact on task completion and factual consistency of the generated responses.

3. A framework for generating and annotating task-oriented document-grounded feedback-annotated dialogue data.

FEDI is comparable to other related datasets in terms of size, lexical diversity and dialogue length (see Table 1). In our analysis, we provide insights into the quality of the generated annotations.

2 Related Work

Methodically Recent work on open-domain dialogue systems shows that considering demographic information has a positive impact on generation accuracy and user engagement in open-domain dialogue systems (Hwang et al., 2023; Lee et al., 2022; Zhang et al., 2018; Siddique et al., 2022; Luo et al., 2019). This is similar for user emotions (Firdaus et al., 2020; Rashkin et al., 2019; Hsu et al., 2018). Using implicit user feedback for this purpose usually requires to train the model with the feedback data (Ung et al., 2022; Xu et al., 2023a; Veron et al., 2021). In this respect, continual learning has shown to be very promising (Xu et al., 2023b; Hancock et al., 2019). This also applies to task-oriented dialogue systems (Wang et al., 2019; Veron et al., 2021; Mazumder et al., 2020). However, as for open-domain dialogue systems, these approaches focus only on the impact on generation accuracy and ignore task completion, which is important to task-oriented dialogue systems (Budzianowski et al., 2018; Nekvinda and Dušek, 2021).

Datasets Table 1 gives a comparison of the datasets resulting from the aforementioned works.

[2]We used the Python package lexical-diversity v0.1.1 for calculation (last accessed 04 January 2024), which implements the approach proposed by McCarthy and Jarvis (2010).

[3]OpenAI GPT-3.5 Model Page (last accessed on 02 January 2024). The model is based on Ouyang et al. (2022). The data was generated between March and June 2023.

For task-oriented dialogues, EmoWOZ (Feng et al., 2022) provides annotations for user emotions, but focuses only on the task of emotion recognition in its experiments. FITS (Xu et al., 2023b) is actually an open-domain dialogue dataset, but provides annotations for knowledge documents (which is why we classify it as document-grounded). However, regarding implicit user feedback, it does not distinguish between different types, e.g., whether the user responds with a correction or asks for clarification, and is limited to generation errors specific to its tasks. This also applies to the other datasets annotated with implicit user feedback (Xu et al., 2023a; Ung et al., 2022). The table also shows that most of the available datasets are the result of crowdsourcing efforts, often leading to datasets of varying quality due to, e.g., methodical artifacts or annotator biases (Yang et al., 2023; Parmar et al., 2023; Thorn Jakobsen et al., 2022; Prabhakaran et al., 2021). As an alternative, recent works suggest synthetic data generation using large language models as a more efficient approach to generate high-quality dialogue data (Kim et al., 2023; Li et al., 2023; Lee et al., 2022), despite their tendency to generate hallucinated or harmful output (Ji et al., 2023; Zhang et al., 2023; Malaviya et al., 2023).

In this work, we generate a dataset to investigate the impact of including implicit user feedback, demographic information and user emotions on task completion and factual consistency of the generated responses in task-oriented document-grounded dialogues. We use the taxonomies provided by Petrak et al. (2023) to cover a variety of generation errors and implicit user feedback types. To address the potential limitations of synthetic data, we recruit human annotators for quality assessment, curation, and to collect a separate set of test dialogues.

3 FEDI

FEDI covers four use cases for task-oriented document-grounded dialogue systems from three domains, including post office services, receptionist services and customer services in the insurance domain. For post office services, we include (1) customer support for parcel shipping, i.e., guiding them through the process of parcel shipping from choosing the right shipping box to informing them about the approximate delivery time, and (2) topping up a prepaid SIM card. For receptionist and customer services in the insurance domain, we include one use case each, i.e., access control (the reception and registration of new visitors in office buildings) and question answering (in the context of financial topics and pet, health and heritage insurance). The question answering dialogues are additionally annotated with the documents that provide the knowledge required for response generation. Appendix A describes the tasks in more detail, including slots, intents, examples, and document sources.

Implicit User Feedback For the generation and annotation of implicit user feedback, we use the user reaction type taxonomy proposed by Petrak et al. (2023), which distinguishes five user reaction types in response to generation errors in preceding system utterances, including Ignore and Continue, Repeat or Rephrase, Make Aware With Correction, Make Aware Without Correction, and Ask for Clarification. For generation errors in system utterances, they also propose an error taxonomy of ten types, nine of which are relevant for task-oriented document-grounded dialogues, such as Ignore Request, Attribute Error, Factually Incorrect, or Lack of Sociality. Definitions, further details and examples can be found in Appendix B.

Demographic Information We consider gender, age, occupation, name, and language style as demographic information in this work. Overall, we distinguish 12 different language styles, such as formal, dialect and jargon, five demographic cohorts, ranging from Boomers (born between 1952 and 1962) to Generation Alpha (born between 2007 and 2016), a variety of $1,155$ occupations, and $2,000$ names. We provide more details, including data sources in Appendix B.

User Emotions We use the taxonomy from EmotionLines (Hsu et al., 2018), which covers seven different emotions, including Neutral, Joy (which we refer to as Happiness), Sadness, Surprise, Fear, Anger, and Disgust. We extend this list with four emotion types we found to be relevant in related work (Kim et al., 2023; Rashkin et al., 2019), including Confusion, Curiosity, Frustration, and Stress. We consider Confusion, Frustration, Fear, Sadness, Disgust, Stress, and Anger as negative emotions.

Problem Formulation We define a dialogue as a set of multiple turns T. Each turn consists of two utterances, a user utterance U_t and a system utterance S_t. Given the dialogue context C =

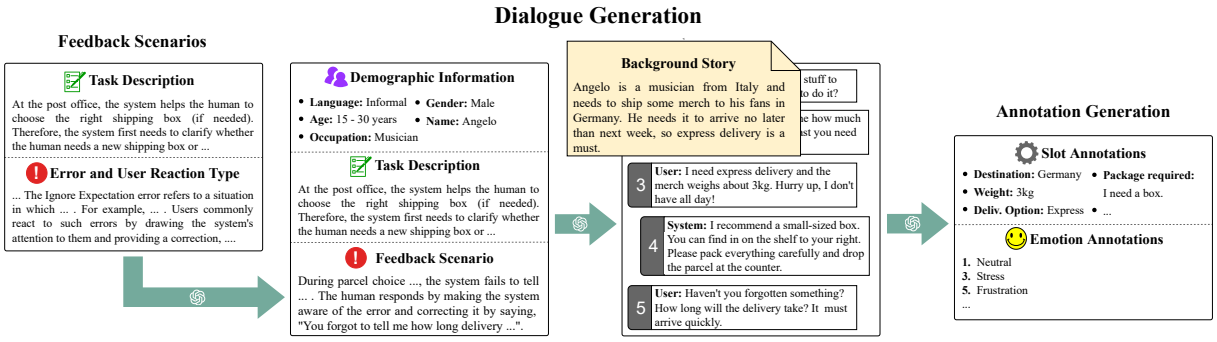

Figure 2: Overview of our framework for generating and annotating dialogues. We distinguish feedback-free and feedback dialogues. The generation of feedback dialogues requires feedback scenarios as additional input.

$[T_0, ..., T_{t-1}]$, and additional information K, the task is to predict the user intent I_t, generate belief state B_t and system utterance S_t:

$$(I_t, B_t, S_t) = \text{generate}(K, C, U_t) \quad (1)$$

Depending on whether knowledge from a document D_t is required to generate S_t or the user emotion E_t, demographic information DI, generation error GE_t, or implicit user feedback F_t should be considered, $K = \{D_t, DI, E_t, GE_t, F_t\}$. DI includes the user's gender, age range, occupation, name, and language style. Belief state B_t includes the slot values inferred from the dialogue context C, which may be used to query knowledge from an external information retrieval system (Chen et al., 2022; Hosseini-Asl et al., 2020), such as the document D_t in the case of customer service or registration information in the case of access control.

4 Framework for Generating and Annotating Dialogues

Figure 2 gives an overview of our framework for generating and annotating dialogues. We distinguish feedback-free and feedback dialogues, i.e., dialogues that provide annotations for implicit user feedback. However, the procedure for dialogue and annotation generation is in general the same for both. For each step that involves GPT-3.5-Turbo, we require the model to return the results in a predefined JSON scheme. If in one step the generation does not match this requirement, the whole dialogue is discarded. We provide more details, including the instructions used in this procedure, in Appendix C.

4.1 General Approach to Dialogue Generation

For dialogue generation, we provide GPT-3.5-Turbo with randomly sampled demographic information for the user, a task description, and the role of the starting actor, i.e., user or system. As indicated by the boxes on the left side of Figure 2, a task description describes the flow of events and information which needs to be conveyed by each role to fulfill the task. In the case of question answering, it also includes a randomly sampled list of documents from the respective topic. Similar to Lee et al. (2022), we instruct the model to use the task description and the demographic information to generate a background story to guide the conversation, such as depicted in the center of Figure 2. We also instruct the model to return the utterance-level annotations for intents (not included in Figure 2) and limit the dialogue to 13 turns, since we found that longer dialogues tend to deviate from the task description. For background stories, we limit the length to five sentences to avoid them becoming a distraction.

Annotation Generation For slot annotations, we provide GPT-3.5-Turbo with the generated dialogue and a list of all slots defined in the task description, possible values and examples[4]. We also instruct the model to only assign and copy values from the dialogue (to prevent hallucinations) and to return the annotations on utterance-level. For emotion annotations, we instruct the model to predict the emotion for each user utterance in the dialogue, given the dialogue and our emotion taxonomy.

[4]We also tried to reduce API calls by combining dialogue and annotation generation, but found that this does not produce reliable results.

4.2 Feedback Dialogues

Feedback Scenarios A feedback scenario describes a generation error and the following implicit user feedback. For generation, we provide GPT-3.5-Turbo with the task description and a list of randomly sampled generation error and user reaction types. To ensure coherence, feedback scenarios must not be mutually exclusive and together form a story in the context of the task description. For each feedback dialogue, we generate three feedback scenarios that are then used as an additional source for dialogue generation (left side of Figure 2)[5].

Feedback Dialogue Generation For feedback dialogue generation, we instruct GPT-3.5-Turbo to consider each feedback scenario in three utterances in the generated dialogue: The system utterance with the generation error, a subsequent user utterance that reflects the user reaction, and a following system utterance that addresses the user reaction. We consider the generated dialogue as Version 1 and generate three additional versions of the same dialogue, each resolving one of the feedback scenarios (Figure 3).

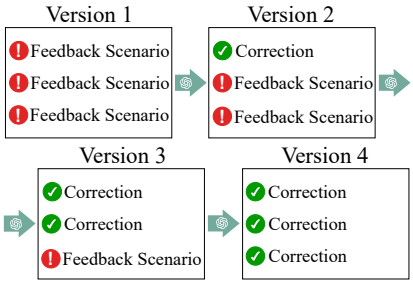

Figure 3: Feedback dialogue generation. Each version resolves one of the feedback scenarios from Version 1.

For each version, we first mask the affected system utterance and generate a replacement using the task description and the preceding dialogue context. Next, we drop the following two utterances, since they are directly related to the generation error. This way, the conversation continues with the next regular user utterance. We continue the process until all feedback turns have been resolved as in Version 4. For slot values, we only regenerate the annotations for the replaced system utterances in Version 2 to 4 and retain the other annotations from Version 1.

[5]We generate all feedback scenarios for a dialogue at once, using the same API call.

5 FEDI Analysis

FEDI consists of 8,852 dialogues, divided into 1,988 feedback-free dialogues, including 326 test dialogues, and 6,864 feedback dialogues (1,716 in four versions, each with one feedback scenario less per dialogue). The test dialogues were collected human-human by eight computer science students in overall 136 paid working hours (see Appendix D for the procedure and details on the hiring process and salary). In the following, we focus on the completeness of generated slot and intent annotations, the distribution of user emotions and the feedback scenarios represented in the dialogues. We provide additional statistical analysis in Appendix E, including split sizes and the distribution of demographic information. In Appendix F, we share our experiences and insights on collecting and annotating dialogue data with humans vs. LLMs.

Slot and Intent Annotations Table 2 shows the ratio of dialogues for which intent and slot annotations were successful, i.e., dialogues that provide all annotations for intent and required slot values.

Task	Feedback-Free Dialogues		Feedback Dialogues			
	Gen.	Test	Version 1	Version 2	Version 3	Version 4
Parcel Shipping	0.87	0.51	0.74	0.72	0.70	0.70
Top Up SIM Card	0.87	0.51	0.74	0.72	0.71	0.69
Access Control	0.86	0.68	0.82	0.83	0.84	0.84
Question Answering	0.99	0.87	0.73	0.99	0.99	0.99

Table 2: The table shows the ratio of dialogues that are complete in the sense that they are annotated with all intent and slot values[6]. For the feedback-free dialogues, we distinguish between generated dialogues (Gen.) and test dialogues.

We observe large differences between (1) question answering and the other tasks and (2) generated dialogues and the test dialogues collected by humans. We found that this is mostly due to variations in the slot annotations. While the slot annotation scheme for question answering is rather simple (see Appendix A), this is different for the other tasks where slots often depend on the background story. For example, in the case of parcel shipping, if the user already has a shipping box and just requires information on the shipping procedure, details about available shipping box types are negligible. While human annotators take this

[6]Hallucinated slot values, i.e. slot annotations that do not occur in the respective utterance, are counted as missing.

into account and occasionally omit slots that are not required based on dialogue motivation, GPT-3.5-Turbo just follows our instructions, which include all slots as part of the task description. For feedback dialogues, we observe that the generated corrections not always address the missing information required by the task description. We provide more analysis as part of our human curation study in Section 6.

Emotion Annotations Figure 4 shows the distribution of the five most common emotions observed in user utterances from both the feedback-free and feedback dialogues[7].

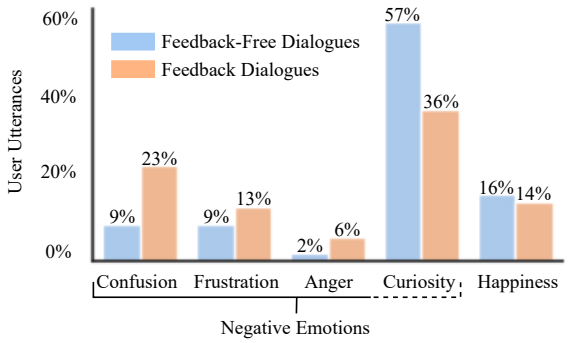

Figure 4: Ratio of the most commonly observed user emotions in FEDI (excluding the Neutral emotion).

As expected, negative emotions are more common in feedback dialogues. For Curiosity, we found that the polarity depends on the dialogue context, e.g. whether the previous system utterance successfully addressed the user's request. It is an emotion that can be either positive or negative, thus it is frequently observed in both dialogues types. Happiness in feedback dialogues is mostly observed in response to system utterances that address user reactions.

Feedback Scenarios Figure 5 shows the distribution of user reactions in relation to error types represented in the feedback scenarios of the feedback dialogues.

The figure shows that our approach for generating feedback scenarios mostly resulted in meaningful combinations of generation error and user reaction types. For example, Factually Incorrect

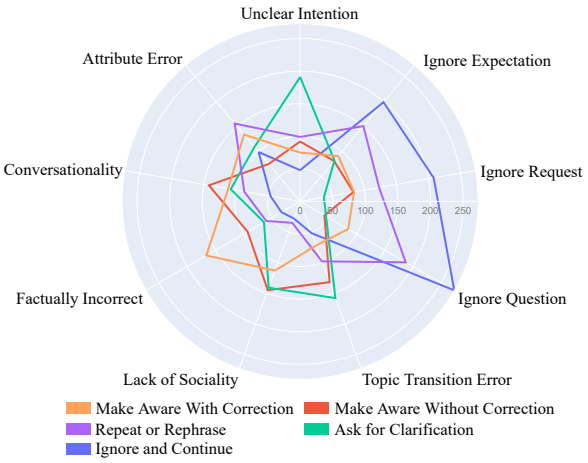

Figure 5: Distribution of user reactions in relation to generation error types represented in feedback scenarios.

is mostly addressed by Make Aware with Correction. Unclear Intention and Attribute Error are frequently addressed by Ask for Clarification and Repeat or Rephrase. The latter one is also frequently observed in combination with Ignore Question and Ignore Expectation errors, although Ignore and Continue is the most frequent user reaction to these generation error types.

6 Human Curation Study

We asked two participants from our test data collection to assess and curate the intent, slot and emotion annotations in 480 feedback-free dialogues and the generation error and user reaction type annotations in 380 feedback dialogues. We used INCEpTION (Klie et al., 2018) as a platform for this study. We calculate the agreement between the annotators using Krippendorff's Alpha (Krippendorff, 2006) as provided in the INCEpTION platform. Table 3 shows the results[8].

	Annotation Type	Missing	Changed	IAA
Feedback-Free Dialogues	Intent	0.06	0.35	0.90
	Slot Values	0.56	0.19	0.83
	User Emotions	0.02	0.81	0.91
Feedback Dialogues	Generation Error Type	0.16	0.36	0.97
	User Reaction Type	0.16	0.34	0.89

Table 3: The ratio of dialogues with at least one missing or changed annotation in our human curation study.

Overall, the ratio of dialogues with at least one missing annotation is rather low, except for slot

[7]We do not distinguish between generated and test dialogues here. We also leave out the neutral emotion as it is in general the most frequently observed emotion (40.5% of all annotated emotions).

[8]Overall, 26 dialogues were reported as off-topic (13 feedback and 13 feedback-free). They are not considered in these results.

annotations. We found that most of them are parcel shipping dialogues, which has a comparatively complex annotation scheme (see Appendix A). We attribute this to our observation on slot and intent annotations from Section 5, i.e., the (occassional) dependence on the background story. A detailed analysis revealed that an average of 1.8 annotations were added to these dialogues. For the dialogues with at least one changed annotation, we found that in many of these cases placeholders, e.g., the slot name put in brackets ([shipping_box_name]), were used instead of the slot values from the dialogues (reported by the students). Emotion is the most frequently changed annotation type (on average 2.09 times per affected dialogue), with the originally-annotated emotion often being Neutral.

7 Experiments and Results

We conduct experiments using three models of different architecture and pretraining approaches, including Flan-T5 (Chung et al., 2022) (780M), GPT-2 (Radford et al., 2019) (780M) and Llama 2 (Touvron et al., 2023b) (7B)[9]. We first finetune the pretrained models to the FEDI scenarios using the feedback-free dialogues and include the demographic information and user emotions as part of the input sequences. For Llama 2, we only finetune the LoRA (Hu et al., 2022) weights in our experiments. We then use the best performing feedback-free models for experiments with the feedback dialogues. Table 4 shows the results achieved in the human-human test dialogues (averaged over three runs). We provide more details, including hyperparameters and input sequences, in Appendix G.

Evaluation Metrics We use F1-Score, BLEU(-n) (Papineni et al., 2002) and BertScore (Zhang et al., 2020) to measure the accuracy of the generated system utterances (generation accuracy). For task completion, we use Inform and Success as proposed by Budzianowski et al. (2018) and measure the correctness of the predicted intents (intent accuracy) and slot values (slot accuracy). To measure the toxicity in the generated responses, we use Perspective API. To measure the factual consistency in the case of question answering, we use Q^2 (Honovich et al., 2021)[10].

Results In general, we find that including demographic information, user emotions and implicit user feedback has a positive impact on task completion and factual consistency of the generated responses, which is particularly important in task-oriented document-grounded dialogues. However, we observe the most significant improvements in the feedback experiments, in which the slot accuracy of Llama 2 (Touvron et al., 2023b) increases by up to 30.9 points, the factual consistency of GPT-2 (Radford et al., 2019) by up to 7.4 points and the intent accuracy of Flan-T5 by up to 28.7 (Chung et al., 2022) points. We attribute these improvements to the additional context provided by the generation error and the user reaction, which can be interpreted as a negative example for a response in the specific dialogue context, but also notice the negative impact on generation accuracy (especially in the case of GPT-2 and Llama 2). We do not observe this in the results on the feedback validation data (Appendix G) and the results of our human evaluation also show that it has no negative impact on user interaction.

Since our feedback dialogues include multiple versions of increasing quality of the same dialogues, it is also possible to use them for continual learning from implicit user feedback. We provide the results in Appendix G.

Human Evaluation We use the two best feedback-trained models from Table 4 (Flan-T5 (Chung et al., 2022) and GPT-2 (Radford et al., 2019) with generation error and user reaction) and their feedback-free counterparts to generate responses for 50 randomly chosen samples from the human-human test dialogues. We then asked two participants from our lab[11] to rate the generated responses for human-likeness (naturalness), relevancy in the dialogue context (coherence), social acceptability (safety), factual consistency (with the target document in the case of question answering), and engagement (whether they would use this model in practice). We use a likert scale from 1

[9]The model weights for Flan-T5 and GPT-2 are available in the Huggingface Model Hub (last accessed 04 January 2024). Access to the weights for Llama 2 must be requested from Meta AI (last acessed 04 January 2024).

[10]We measure the F1-Score based on the overlapping tokens in target and prediction. For BLEU and BertScore, we use the implementation from the HuggingFace evaluation library v0.4.1 (last accessed 04 January 2024) and with $n = 4$ for BLEU. For Inform and Success, we use the implementation from Nekvinda and Dušek (2021) as a reference (last accessed 04 January 2024). For Q^2, we use the reference implementation which is available in GitHub (last accessed 04 January 2024). Perspective API is a free-to-use service provided by Google and Jigsaw. Model and training details can be found here (last accessed 04 January 2024).

[11]One of the authors and an intern of our research group who participated during their working hours.

	Experiment	Generation Accuracy			Task Completion				Quality	
		F1	BLEU	BertScore	Inform	Success	Intent Acc.	Slot Acc.	Toxicity	Q^2
Flan-T5 Feedback-Free	Flan-T5	45.0	20.0	88.3	86.7	85.9	54.8	60.9	0.02	52.7
	+Emotions	**46.7** (+1.7)	**21.0** (+1.0)	**88.9** (+0.6)	**83.9** (–2.8)	**83.2** (–2.7)	**61.2** (+6.4)	**58.3** (–2.6)	**0.02**	**57.5** (+4.8)
	+Demographics	43.2 (–1.8)	18.4 (–1.6)	87.7 (–0.6)	87.0 (+0.3)	86.0 (+0.1)	33.5 (–21.3)	29.3 (–31.6)	0.03 (+0.01)	54.5 (+1.8)
	+Emotions +Demographics	44.2 (–0.8)	19.1 (–0.9)	88.1 (–0.2)	85.3 (–1.4)	85.1 (–0.8)	43.9 (–10.9)	36.7 (–24.2)	0.02	56.4 (+3.7)
Feedback	+Generation Error	41.4 (–3.6)	19.8 (–0.2)	87.8 (–0.5)	96.8 (+10.1)	92.7 (+6.8)	72.5 (+17.7)	76.7 (+15.8)	0.02	56.9 (+4.2)
	+User Reaction	41.3 (–3.7)	19.3 (–0.7)	87.6 (–0.7)	96.6 (+9.9)	94.1 (+8.2)	69.0 (+14.2)	76.2 (+15.3)	0.02	56.3 (+3.6)
	+Generation Error +User Reaction	**44.4** (–0.6)	**22.1** (+2.1)	**88.2** (–0.1)	**96.9** (+10.2)	**95.3** (+9.4)	**83.5** (+28.7)	**77.2** (+16.3)	**0.02**	**60.2** (+7.5)
GPT-2 Feedback-Free	GPT-2	34.9	10.4	87.1	88.3	81.6	78.7	69.6	0.02	28.1
	+Emotions	35.1 (+0.2)	10.4	87.1	84.1 (–4.2)	83.8 (+2.2)	75.4 (–3.3)	67.3 (–2.3)	0.02	26.7 (–1.4)
	+Demographics	34.6 (–0.3)	10.4	87.1	80.2 (–8.1)	80.2 (–1.4)	69.3 (–9.4)	57.5 (–12.1)	0.02	26.3 (–1.8)
	+Emotions +Demographics	**36.0** (+1.1)	**11.4** (+1.0)	**87.3** (+0.2)	**85.1** (–3.2)	**84.8** (+3.2)	**71.6** (–7.1)	**66.7** (– 2.9)	**0.02**	**29.2** (+1.1)
Feedback	+Generation Error	29.2 (–5.7)	8.0 (–2.4)	86.2 (–0.9)	92.4 (+4.1)	91.7 (+10.1)	84.3 (+5.6)	79.3 (–9.7)	0.02	30.9 (+2.8)
	+User Reaction	30.0 (–4.9)	8.3 (–2.1)	86.3 (–0.8)	98.9 (+10.6)	96.5 (+14.9)	83.0 (+4.3)	80.3 (+10.7)	0.02	32.3 (+4.2)
	+Generation Error +User Reaction	**30.3** (–4.6)	**9.7** (–0.7)	**86.4** (–0.7)	**94.7** (+6.4)	**93.3** (+11.7)	**88.0** (+9.3)	**80.8** (+11.2)	**0.01** (–0.01)	**35.5** (+7.4)
Llama 2 Feedback-Free	Llama 2	29.3	7.1	86.1	85.9	81.2	37.6	39.2	0.02	28.3
	+Emotions	**36.3** (+7.0)	**14.9** (+7.8)	**85.4** (–0.7)	**89.3** (+3.4)	**85.3** (+4.1)	**40.2** (+2.6)	**41.3** (+2.1)	**0.01** (–0.01)	**18.7** (–9.6)
	+Demographics	33.8 (+4.5)	4.5 (–2.6)	86.5 (+0.4)	85.6 (–0.3)	82.5 (+1.3)	37.1 (–0.5)	40.1 (+0.9)	0.02	21.3 (–7.0)
	+Emotions +Demographics	28.8 (–0.5)	5.6 (–1.5)	81.3 (–4.8)	86.7 (+0.8)	87.9 (+6.7)	41.4 (+3.8)	39.6 (+0.4)	0.03 (+0.01)	20.6 (–7.7)
Feedback	+Generation Error	24.1 (–5.2)	7.9 (+0.8)	77.4 (–8.7)	93.1 (+7.2)	95.7 (+14.5)	54.8 (+17.2)	59.6 (+20.4)	0.01 (–0.01)	29.1 (+0.8)
	+User Reaction	**24.5** (–4.8)	**6.9** (–0.2)	**78.8** (–7.3)	**94.9** (+9.0)	**93.2** (+12.0)	**63.5** (+25.9)	**70.1** (+30.9)	**0.02**	**27.1** (–1.2)
	+Generation Error +User Reaction	25.0 (–4.3)	9.2 (+2.1)	80.1 (–6.0)	82.4 (–3.5)	83.6 (+2.4)	46.3 (+8.7)	47.2 (+8.0)	0.03 (+0.01)	33.5 (+5.2)

Table 4: Results of our experiments. We use the pretrained models finetuned on the feedback-free dialogues (Feedback-Free) as deltas. The best performing models are highlighted and, in the case of feedback-free experiments, are used for the experiments with feedback dialogues (Feedback). Learning from user emotions (+Emotions) has a positive impact on the generation accuracy. The demographic information (+Demographics) is of minor importance. Learning from implicit user feedback (+User Reaction) and the preceding generation error (+Generation Error) leads to improvements in terms of task completion and factual consistency of the generated responses (Q^2).

(lowest rating) to 5 (highest rating) for each attribute and provide the annotators with the knowledge document, dialogue context, and generated response for this evaluation. The order of the dialogues was randomized to prevent the annotators from drawing conclusions about the generating model. Table 5 shows the results.

Experiment	Naturalness	Coherence	Safety	Engagement	Factual Consistency
Flan-T5					
Feedback-Free	4.14	4.15	4.55	3.75	2.20
Feedback	4.30 (+0.16)	4.25 (+0.10)	4.59 (+0.04)	3.89 (+0.14)	2.24 (+0.04)
GPT-2					
Feedback-Free	4.24	3.82	4.45	3.44	1.55
Feedback	4.42 (+0.18)	4.05 (+0.23)	4.46 (+0.01)	3.76 (+0.32)	1.58 (+0.03)

Table 5: Results of our human evaluation.

In general, the responses generated by the feedback-trained models are rated higher, although the differences are rather marginal. The participants also reported that these responses encourage for more user interaction, e.g., by requesting additional information or paying more attention to the user and their situation, and are in general more factual consistent. In the case of question answering, the generated responses are mostly summaries of the respective documents. According to the annotators, the GPT-2 model trained only with emotions already produced very engaging answers, although they were not as coherent as the responses generated by Flan-T5. This also affects factual consistency in the case of question answering.

8 Conclusion

In this work, we investigated the impact of learning from implicit user feedback, demographic information and user emotions on task completion and factual consistency of the generated responses in task-oriented document-grounded dialogues. We also introduced FEDI, the first English dialogue dataset that provides annotations for such data. Our analysis shows the effectiveness of our generation framework and that FEDI is comparable to other related datasets. Our experiments with Flan-T5, GPT-2 and Llama 2 show that including implicit user feedback, demographic information and user emotions has a positive impact on task completion and factual consistency. For future work, we are planning to improve the quality of annotations in FEDI and to generalize our generation framework to other tasks.

9 Limitations

The training and validation dialogues in FEDI were synthetically generated using GPT-3.5-Turbo. Thus, there is a probability that some data is unfaithful, hallucinated, or even harmful (Kumar et al., 2023; Zhang et al., 2023; Malaviya et al., 2023). In addition, some of these dialogues may seem artificial and unnatural due to potentially conflicting demographic information, e.g., language style contradicting with age or occupation. The same applies to the feedback scenarios represented in the feedback dialogues. It is possible that some user reactions appear to be unnatural, counterintuitive, and maybe not even addressing the underlying generation error. However, we did not observe any of these issues in our analysis.

In our experiments, we analyzed the toxicity of generated responses using the Perspective API. We acknowledge that the detector may not capture all the potentially harmful content. The generated data may also contain positive stereotypes, i.e., seemingly-harmless words or patterns that are offensive to specific demographic groups, which are not marked by the detector (Cheng et al., 2023). The final human evaluation was conducted with only two people from our research group due to a lack of resources. One of them was one of the authors. Although we have tried to preclude the possibility that the annotators can draw conclusions about the generating model, this limits the significance of the results. However, the fact that the results between the feedback-annotated and feedback-free models are close (although the feedback-annotated models are slightly better) suggests that the evaluation was carried out fairly.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple

language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 2006. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6794–6801. AAAI Press.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *CoRR*, abs/2309.07852.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Sepideh Esmaeilpour. 2020. An application-independent approach to building task-oriented chatbots with interactive continual learning. In *NeurIPS-2020 Workshop on Human in the Loop Dialogue Systems*.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Wei Peng Minjin Rheu, Ji Youn Shin and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction*, 37(1):81–96.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang,

Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.

Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855.

Dominic Petrak, Nafise Moosavi, Ye Tian, Nikolai Rozanov, and Iryna Gurevych. 2023. Learning from free-text human feedback – collect new datasets or extend existing ones? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16259–16279, Singapore. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

A.B. Siddique, M.H. Maqbool, Kshitija Taywade, and Hassan Foroosh. 2022. Personalizing task-oriented dialog systems via zero-shot generalizable reward function. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1787–1797, New York, NY, USA. Association for Computing Machinery.

Terne Sasha Thorn Jakobsen, Maria Barrett, Anders Søgaard, and David Lassen. 2022. The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 44–61, Marseille, France. European Language Resources Association.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeR-Dialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.

Mathilde Veron, Sophie Rosset, Olivier Galibert, and Guillaume Bernard. 2021. Evaluate on-the-job learning dialogue systems and a case study for natural language understanding. *CoRR*, abs/2102.13589.

Weikang Wang, Jiajun Zhang, Qian Li, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. 2019. Incremental learning from scratch for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3710–3720, Florence, Italy. Association for Computational Linguistics.

Jing Xu, Da Ju, Joshua Lane, Mojtaba Komeili, Eric Michael Smith, Megan Ung, Morteza Behrooz, William Ngan, Rashel Moritz, Sainbayar Sukhbaatar, Y-Lan Boureau, Jason Weston, and Kurt Shuster. 2023a. Improving open language models by learning from organic interactions. *CoRR*, abs/2306.04707.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023b. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

A Task Descriptions

In the following, we provide details on the tasks included in FEDI and their slot values. Following (Budzianowski et al., 2018), we distinguish requestable and informable slots, since this is necessary to calculate the task completion metrics in Section 7.

Post Office Services FEDI includes dialogues from two basic services provided in post offices, customer support for parcel shipping and topping up a prepaid SIM card. In customer support for parcel shipping, the task is to help the user choose the right shipping box and delivery option for their needs (given the weight of the goods to be sent and the destination). Topping up a prepaid SIM card is less of an advisory service since customers usually know how much they want to recharge, their telephone number, and which telephone provider they are with. Table 6 lists the slots for each task.

Slot Name	Informable	Requestable	Description
Parcel Shipping			
Destination	✓		The city and country of destination; national or international.
Weight	✓		The weight of the item to be shipped, lightweight (up to 5kg), average (up to 20kg), heavy (up to 30kg).
Package Required	✓		Whether or not a new shipping box is required.
Delivery Option	✓		Express or standard delivery.
Country of Destination	✓		The destination country.
Shipping Box Name		✓	Name of the best suitable shipping box (small-sized, medium-sized, large-sized), based on the weight of the item to be sent.
Shipping Box Description		✓	Brief description on why the suggested shipping box is a good choice.
Shipping Procedure		✓	Description of the shipping procedure (e.g., take the box to the counter...)
Shipping Time		✓	Expected delivery time, one to three days for national, four to six days for european, and 3-4 weeks for international deliveries.
Top Up SIM Card			
Phone Number	✓		Table or mobile phone number with country code, e.g., +39 XXX XXXXXXX.
Phone Provider	✓		The phone provider, e.g. Vodafone, POSTE Mobile, ...
Import Payment	✓		The recharge amount, e.g., 10 euro, 20 euro, 30 euro.
Outcome Operation		✓	If all required information were provided, the system asks the user to insert the card for payment.
Request Ticket			
Type of Service	✓		The type of service for which the user wants to request support, i.e., parcel shipping or top up prepaid SIM card.
Ticket Number		✓	The ticket number generated for the request.

Table 6: Slot values for parcel shipping and topping up a prepaid SIM card.

In modern post offices, service robots or other virtual agents are more commonly used to provide such services in a self-service manner. However, if something goes wrong, e.g., the shipping boxes are empty or the credit card was rejected, customers must have the option of requesting assistance from a human employee. In this case, the customer is asked to tell the agent the type of service they need assistance with. In turn, the agent creates a ticket for a human employee and returns the ticket number. We consider this as a kind of subtask to the other tasks (Request Ticket in Table 6) and do not evaluate it separately.

Receptionist Services For receptionist services, FEDI only includes one task: access control. Table 7 shows the slots for this task.

Slot Name	Informable	Requestable	Description
Access Control			
Guest Name	✓		The name of the person who wants to access the building.
Host Name	✓		The name of the person the guest wants to visit.
Host E-Mail	✓		The E-Mail address of the host.
Alternative Host Name	✓		An alternative host, e.g., in case the host is not available.
Alternative Host E-Mail	✓		E-Mail address of the alternative host.
Meeting Date and Time	✓		Date and time of the appointment.
Meeting Room Identifier	✓		Unique identifier of the room where the meeting will take place.
Verification Call		✓	The system can set up a verification call to let the host visually inspect the guest and authorize access.
Confirmation to Open Turnstile		✓	This is a signal to the system that controls the turnstile to let the guest enter.
Add. Safety Information		✓	Any additional safety information, e.g., related to COVID-19.

Table 7: Slot values for access control.

It is an essential task in hotels, office buildings, or other facilities with restricted access. Visitors usually need to register at the reception desk before being allowed to enter. As of today, electronic access controls (EAC) are more common than reception desks, especially in the case of office buildings, and they are becoming increasingly intelligent. In our case, we focus on a scenario in which a visitor has an appointment with an employee in an office building. To access the building, the visitor needs to provide the EAC with information about the appointment, e.g., the name of the host, date and time, and the room number. The EAC can then decide to grant access or to call the host for confirming the visitor's identity. If necessary, the EAC can also provide additional safety information, e.g., hygiene guidelines.

Customer Service in the Insurance Domain For customer service in the insurance domain, we focus on question answering in the context of pet, health or heritage insurance, as well as bank transactions and account conditions. As a source, we use the insurance policies from POSTE Italiane, which are also available in English language[12]. Table 8 lists the slots.

Slot Name	Informable	Requestable	Description
Question Answering			
Question	✓		A question related to one of the topics.
Type of Bills	✓		If the user asks a question regarding a specific payment slip, they need to provide the type.
Evidence		✓	The answer to the user's question.
Bill Form Description		✓	Description of the specific payment form (if the question was about a payment form).
Bill Form Name		✓	Name of the payment form (if the question was about a payment form).
Bill Form Payment Procedure		✓	Information on how to fill the payment form (if the question was about a payment form).

Table 8: Slot values for question answering.

In the past, customers called their insurance agent or visited their local bank branch for all questions related such topics. Today, it is more common to talk to chatbots or other service agents first and only in exceptional cases to human employees. Overall, we extracted 313 question-document pairs, i.e., questions paired with a paragraph that contains the answer, 19 for bank transactions, 93 for account conditions, 78 for health, 84 for heritage, and 39 for pet insurance, from the POSTE documents.

Greeting In the prompts for dialogue generation (see Appendix C), we instruct GPT-3.5-Turbo to have a separate turn at the beginning and ending of a dialogue in which both roles greet each other by also considering the generated background story. However, we do not consider this as a separate task in the sense of this work and do not evaluate it separately.

B Dataset Features

In this section, we provide additional details on the demographic information and the error and user reaction types used to create FEDI.

Demographic Information We distinguish 12 different language styles, including Their Age and Job, Standard, Colloquial, Formal, Gutter, Polite, Informal, Regional Dialect, Social Dialect, Jargon, Slang, and Age. For age ranges, we consider five

[12] POSTE Italiane Insurance Policies, last accessed 13 January 2024.

demographic cohorts, including Boomers (born between 1952 and 1962), Generation X (born between 1962 and 1977), Millenials (born between 1977 and 1992), Generation Z (born between 1992 and 2007), and Generation Alpha (born between 2007 and 2016). For occupations, we use a list of 1,155 job titles sampled from The Gazette[13], including among others jobs from the fields of science and technology, education, arts and entertainment, healthcare, or manufacturing. As a source for the names, we use the list of the 2,000 most popular American baby names in 2010[14]. For each dialogue, we randomly sample a new value for each characteristic and apply simple plausibility checks, e.g., a person from Generation Alpha can only be a pupil.

Error and User Reaction Types To generate implicit user feedback, we use the generation error and user reaction type taxonomy proposed by Petrak et al. (2023). For generation errors in system utterances they define the following nine error types as relevant for task-oriented document-grounded dialogues:

- **Ignore Question** — This error occurs when the system fails to address a user's question. Instead of providing a relevant response or clarification, the system disregards their input.
- **Ignore Request** — A situation in which the system fails to take action on a user's request. It can occur due to various reasons, such as misinterpretation of the request, technical limitations, or system glitches.
- **Ignore Expectation** — This error happens when the system fails to fulfill the user's expectation in terms of understanding and addressing their needs within the context of the task.
- **Attribute Error** — If the system fails to correctly extract or understand the necessary slots or attributes from a user's utterance, this is called an attribute error.
- **Factually Incorrect** — System responses that are factually wrong or inaccurate.
- **Topic Transition Error** — A situation in which the system's response abruptly shifts to a different or previously discussed topic without a logical connection or adequate context.
- **Conversationality** — Bad conversationality occurs when the system fails to maintain a coherent and natural conversation flow, e.g., it repeats previous responses or contradicts itself without recognizing or asking for new or missing information.
- **Unclear Intention** — This error is characterized by the system's failure to accurately address a user's intended objective.
- **Lack of Sociality** — If a system's response doesn't adhere to social conventions, fails to include basic greetings, or exhibit toxic and disrespectful behavior or language, this is referred to as a lack of sociality.

They also define an error type for common sense errors, but found them to be rare in task-oriented document-grounded dialogues. For this reason, we do not consider this error type in our work.

For user reactions in response to generation errors, they propose the following taxonomy:

- **Ignore and Continue** — The user ignores the error and continues the conversation, e.g., "Okay. Let's leave it like that.".
- **Repeat or Rephrase** — Instead of ignoring the error in the system utterance, the user repeats or rephrases their original concern, e.g., "Actually, I wanted you to ...".
- **Make Aware With Correction** — The user makes the system aware of its error and provides a correction or response alternative, e.g., "Partly. This doesn't take into account that ...".
- **Make Aware Without Correction** — Instead of providing a correction or response alternative, the user just makes the system aware of its error, e.g., "You're wrong.".
- **Ask for Clarification** — In case of error, the user asks the system for clarification, e.g., "I'm not sure what you mean. Is it about ...".

C Prompts for Dialogue Generation and Annotation

In this section, we provide more details on the steps and prompts used for generating FEDI. Addition-

[13] Available in GitHub (last accessed on 16 January 2024).

[14] Published by babymed.com (last accessed 12 February 2024).

ally added source data is highlighted in blue in the figures below.

JSON Schemes As described in Section 4, we require GPT-3.5-Turbo to return all results in a predefined JSON scheme, which depends on the prompt, i.e., dialogue generation or annotation, and ensures that the returned values contain all required fields and is processable without human intervention. If the values returned do not adhere to the required scheme, we drop the whole dialogue. Figure 6 shows an example for the annotation of emotions.

Provide your results in machine-readable json format (escape " and avoid non utf-8 characters). Here is an example:

```
{
  "result": [
    "happiness",
  ]
}
```

Figure 6: Instruction to return the results in json for emotion annotation.

We append these json schemes at the end of the prompts. We basically provide the required fields and example values, and instruct the model to return only utf-8 encoded characters and escape quotation marks (so that we can treat it as a string in Python). Please refer to our GitHub repository for all prompts and their json schemes[1].

Feedback-Free Dialogues For dialogue generation, we distinguish feedback-free and feedback dialogues. Figure 7 shows the instruction used to generate feedback-free dialogues.

Generate a dialogue (max. 13 turns) between a human and a dialogue system in the following task: {name of the task}. For the human, imagine a person ({occupation}, between {age} years old) called {name} that uses {language} language style with a short emotional and task-related background story of max. 5 sentences (including the human's country of residence). Generate the dialogue in a role-play manner. The dialogue system is empathetic and replies and interacts with the human according to their persona and background story. Do not include personal information (e.g., the person's name) in the dialogue. The {role of the starting actor} starts. The conversation begins and ends with a greeting.
{task description}
For each utterance, include the intent (the task addressed) in the json output.

Figure 7: Instruction for generating feedback-free dialogues.

We provide GPT-3.5-Turbo with the demographic information, the role of the starting actor, and the task description. We require the model to use this information to generate a background story and to use this as an additional source for dialogue generation. We also instruct the model to return the utterance-level annotations for intents in this step.

{names of error types} are common generation errors in dialogues.
{list of error type definitions}
Users commonly react to such errors by {user reaction types}. Combine each of these user reaction types with an error type. Then generate a feedback scenario (up to 4 sentences, including why and how it reflects the respective error type) for 3 of these combinations in the following task:
{task description}
It is important that the feedback scenarios are different but not mutually exclusive and together make a story. For each feedback scenario, provide a precise description as continuous text (no dialogues), including the user's reaction and why and how the scenario reflects the respective generation error.

Figure 8: Instruction for generating the feedback scenarios.

Feedback Dialogues Figure 8 shows the instruction for the generation of feedback scenarios, which are required as an additional source for feedback dialogues. We generate them in a separate step before dialogue generation. We generate three feedback scenarios using the same prompt in a separate step before dialogue generation. Figure 9 shows the instruction for the generation of feedback dialogues.

The instruction is longer and more detailed than the one used for generating the feedback-free dialogues (Figure 7). For example, it explicitely describes how to process feedback scenarios. Another difference is the length limitation. While feedback-free dialogues are restricted to 13 turns, we require feedback dialogues to have at least 13 turns. In practice, the length of the feedback dialogues is similar to the length of the feedback-free dialogues, but we observed that feedback dialogues are likely to be cut off without this requirement. We consider the generated dialogue as Version 1 and generate three additional versions of the same dialogue, each resolving one of the feedback scenarios. For this, we regenerate the system utterance with the generation error. Figure 10 shows the instruction.

The instruction includes the dialogue up to the next generation, the name of the task and the respective document (in the case of question answering). We mask the system utterance with generation error using <mask>, since we found this to produce replacements more coherent to the dialogue history.

Generate an erroneous long and in-depth dialogue (at least 13 utterances) between a human and a dialogue system. For the human, imagine a person ({occupation}, between {age} years old) called {name} that uses {language} language style with a short emotional and task-related background story of max. 5 sentences (including the human's country of residence). Generate the dialogue in a role-play manner. Play the dialogue system as not helpful and inattentive. Do not include personal information (e.g., the person's name) in the dialogue. The {role of the starting actor} starts. The conversation begins and ends with a greeting.
{task description}
A feedback scenario consists of a system utterance, in which the dialogue system makes an erroneous statement, and a subsequent human utterance, in which the human reacts to the error in the system utterance in the predefined way. Next, the system responds considering the reaction of the person. Then the situation is done. Generate the dialogue using the following {number} feedback scenarios (all must be included):
{feedback scenarios}
Highlight the erroneous system utterance by adding the respective scenario identifier to the error field of the utterance and to the error field of the following person utterance. Errors always originate from system utterances. Each scenario can only occur twice, once in a system utterance and once in the subsequent human utterance.

Figure 9: Instruction for generating feedback dialogues.

Given is the following turn-based {name of the task} dialogue between a human and a dialogue system. One system utterance is masked using the <mask> token.
{dialogue}
Predict the next system response (max. 4 sentences), using the following information:
{document}
The dialogue system is an empathetic and friendly virtual assistant.

Figure 10: Instruction to regenerate the system utterance to replace the one with the generation error. The document is only included in case of question answering.

After replacing the affected system utterance, we regenerate its slot values. We remove the following two utterances to ensure that the dialogue flow is not corrupted (since they directly refer to the generation error). The conversation then continues with the next regular user utterance. This solution is the result of multiple experiments with different approaches:

- Using the implicit user feedback and the task description and instruct GPT-3.5-Turbo to rewrite the whole dialogue.
- Providing GPT-3.5-Turbo with the whole dialogue and only instruct it to rewrite the affected turn.
- Using the respective feedback scenario as additional input to regenerate the affected system utterance.

All of them resulted in inconsistent dialogues and off-topic, unnatural or incorrect system utterances. During prompt engineering, we found that the feedback itself is negligible for resolving feedback scenarios using our prompt from Figure 10, since it includes the dialogue context and requires the new system utterance to be generated in a friendly and polite manner.

Slot Annotations Figure 11 shows our instruction for generating slot values.

Given is the following dialogue between a dialogue system and a person:
{dialogue}
Identify and copy the corresponding sequences for each of the following slots in the person utterances: {list of slots in person utterances with examples}. Identify and copy the corresponding sequences for each of the following slots in the system utterances: {list of slots in system utterances with examples}.

Figure 11: Instruction for slot annotation in a generated dialogue.

For this, we provide GPT-3.5-Turbo with the complete dialogue and distinguish between slots for each role (person and system). The slots to be annotated are provided in lists (including example values). We also instruct the model to just use sequences from the dialogue as slot values (to avoid hallucinated slot values).

Emotion Annotations Figure 12 shows the instruction for emotion generation. We generate emotions just based on the dialogue context. We do not provide additional information, such as examples. However, we additionally provide the number of utterances in the dialogue and those related to the user.

Given is the following dialogue between a dialogue system and a person (user):
{dialog}
The dialog consists of {number of utterances} utterances, {number of person utterances} of which are person utterances. For each of the person utterances, predict the underlying emotion. This is the list of possible emotions: anger, confusion, curious, disgust, fear, frustration, happiness, neutral, sadness, stressed, surprise.

Figure 12: Instruction for generating emotions.

D Studies With Human Annotators

In this work, we conducted several studies that required human annotators: (1) the test data collection for FEDI, (2) the curation and assessment of the generated annotations and (3) the human evaluation of our trained models. While the human evaluation was conducted by one of the authors and an intern from our research group (who participated during their working hours), the test data collection and curation study was conducted with external participants who were hired for this purpose.

Application Criteria and Hiring Procedure For participation, we required a formal application. Our criteria were as follows:

- Enrollment in computational linguistics, linguistics, data and discourse studies, computer science, business informatics or comparable.
- Fluent in reading, speaking and writing English.
- Good communication and organization skills.

We considered a background in NLP, interest in conversational AI and experience in data annotation as a plus. We did not restrict the job advertisment to our university. Also, we did not take gender into consideration. We asked all applicants that fulfilled those criteria to take part in a recruitment test, in which we asked them to collect and annotate dialogues in a self-chat manner, given a task description from our work. We then assessed and ranked their results based on (1) time needed for one dialogue, (2) annotation completness, (3) number of turns per dialogue, (4) avg. utterance length.

Overall, we received 11 applications that fulfilled our criteria. Eight of them passed the recruitment test and were hired for an hourly salary of 12,95$. While all participants took part in the test data collection only two were involved in the data curation study.

Test Data Collection For test data collection, we randomly assigned participants to groups of two to collect the dialogues in one hour sessions dedicated to one task. For each task, we provided the task description, including slots with examples and four persona profiles (combinations of demographic information) and background stories as inspiration. However, we encouraged them to think about own persona profiles and background stories. For user emotions, we provided them with a list of available options. For question answering, we provided them with the question-document pairs extracted from the POSTE Italiane data (Section A).

For data collection, we use a self-developed web-based platform that allows to collect dialogues between two humans or between a human and a language generation model. Figure 13 shows the user interface.

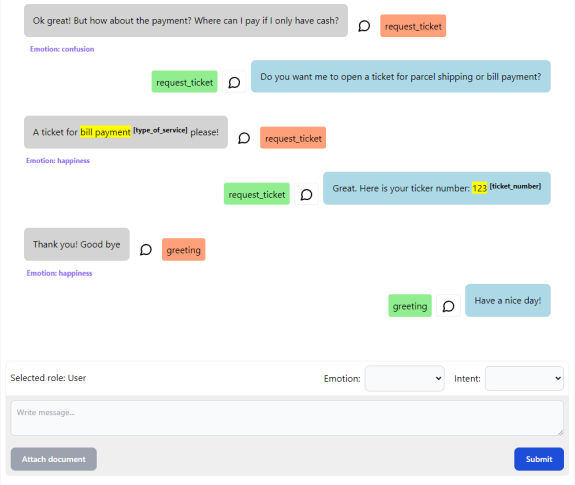

Figure 13: The user interface of the data collection platform used to collect the test data.

Each message is annotated with the respective intent (orange or green, depending on the role). Slot annotations are highlighted in yellow, with the slot type as superscript. Although we did not use it in this study, the chat pane provides the possibility to annotate errors and implicit user feedback by clicking on 🗩 (the speech bubble button between the messages and the intent annotations). For Question Answering, the chatpane also offers the possibility to attach a document to a message (a text file).

E FEDI– Additional Analysis

In this section, we provide additional analysis about the composition of FEDI. Overall, FEDI consists of 8,852 dialogues, 1,988 feedback-free and 6,864 feedback dialogues. Table 9 shows the distribution of dialogues in the dataset. Test for the feedback-free dialogues refers to the human-human collected test data.

Demographic Information Figure 14 shows the distribution of language styles, age ranges and occupations randomly sampled for background story generation.

Task	Feedback-Free Dialogues			Feedback Dialogues				
	Train	Dev	Test	Version 1	Version 2	Version 3	Version 4	Dev
Parcel Shipping	186	20	38	193	193	193	193	84
Top Up SIM Card	187	20	39	193	193	193	193	84
Access Control	183	20	42	215	215	215	215	92
Question Answering	943	103	207	945	945	945	945	420
Per Split	**1,499**	**163**	**326**	**1,546**	**1,546**	**1,546**	**1,546**	**680**
Total			**1,988**					**6,864**

Table 9: Data splits included in FEDI and their sizes.

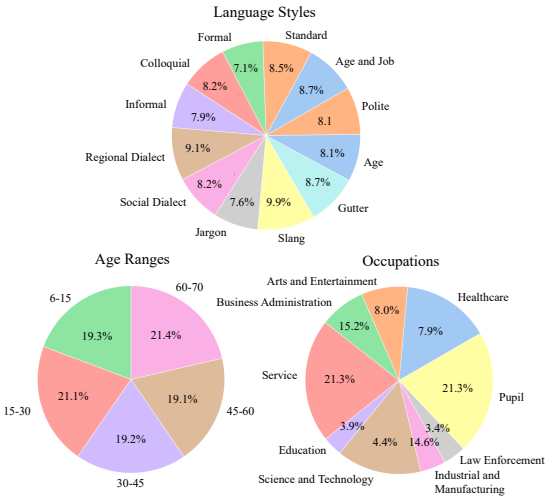

Figure 14: The distribution of persona attributes represented in the background stories (excluding human-human test dialogues).

Language styles are almost equally weighted. For occupations, the figure shows that jobs from the categories of business administration, service, industrial and manufacturing, and pupil largely outweigh the other categories, which makes sense in the context of the tasks and topics represented in FEDI[15]. Overall, we observe 693 unique job titles in FEDI. The figures do not show the distribution of names. We found 1,496 different names in the dialogues. 638 (42%) are unique, and 712 (47.59%) occur two to three times. The remaining 146 names occur four or more times throughout the entire dataset.

Emotions The chart in Figure 15 shows the distribution of emotions in the dialogues of FEDI.

With 40.5%, Neutral is the most common emotion, followed by Curiosity (27.5%). Frustration and Confusion are relatively rare. We observe them mostly in the feedback dialogues. Others refers to emotions that are represented $\leq$ 5%, including

[15]The original list did not provide categories. We generated them using GPT-3.5-Turbo.

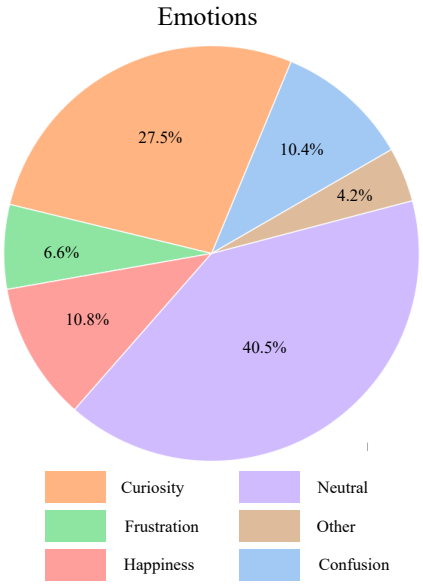

Figure 15: Illustration of the distribution of emotions in FEDI.

Anger, Disgust, Fear, Surprise, and Stress.

Feedback Scenarios Overall, we generated 4,714 feedback scenarios that are included in the 1,716 feedback dialogues of Version 1. Figure 16 shows the distribution of generation error and user reaction types.

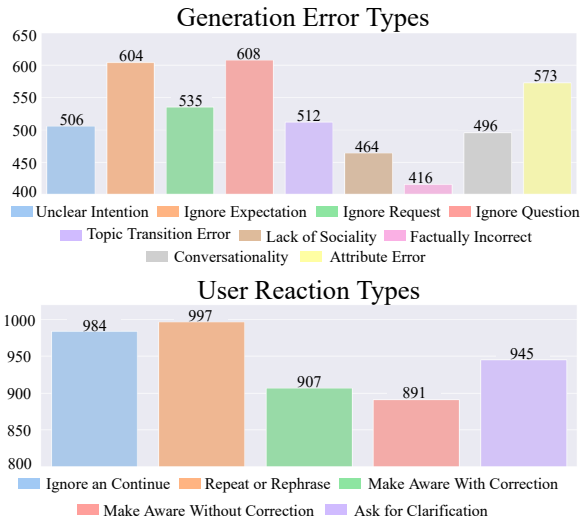

Figure 16: Distribution of generation error and user reaction types in the feedback dialogues of FEDI.

Given that most of the dialogues are about question answering (Table 9), it is not surprising that Ignore Question is the most frequent error type. Table 10 shows the ten most commonly observed error and user reaction type combinations.

Ignore Question and Ignore Request are two of

	Error Type	Feedback Type	Frequency
1	Ignore Question	Ignore and Continue	273
2	Ignore Request	Ignore and Continue	208
3	Ignore Expectation	Ignore and Continue	199
4	Unclear Intention	Ask for Clarification	191
5	Ignore Question	Repeat or Rephrase	187
6	Factually Incorrect	Make Aware With Correction	166
7	Topic Transition Error	Ask for Clarification	158
8	Attribute Error	Repeat or Rephrase	156
9	Ignore Expectation	Repeat or Rephrase	151
10	Lack of Sociality	Make Aware Without Correction	141

Table 10: The table shows the most common error and user reaction type combinations included in FEDI.

the most frequent error types. While we observe the first one more common in question answering dialogues, the second one is more common in the other tasks. For both we observe that Ignore and Continue is the most frequent user reaction type, followed by Repeat or Rephrase. Unclear Intention is an error type mostly observed in parcel shipping, topping up a prepaid SIM card, and access control. The most frequently observed user reaction to this is Ask for Clarification. Based on absolute numbers, Factually Incorrect is the rarest error type. It is mostly observed in question answering and in combination with Make Aware With Correction.

F Dialogue Collection: Human vs. LLM

In our human data collection, eight students collected 326 test dialogues in 136 paid working hours. With an hourly salary of 12.95$, this adds up to a cost of 1,761.20$ (not including additional costs, such as for supervision). Generating and annotating 8,526 dialogues using GPT-3.5-Turbo cost 75.73$, including API calls for prompt engineering and debugging. On average, collecting and annotating a human-human dialogue cost 5.40$. Using GPT-3.5-Turbo, it is 0.009$. Based on this, collecting and annotating dialogues with human participants is rather uneconomic and inefficient. However, with 175B parameters, GPT-3.5-Turbo is an extremely large model. Without access to such a model, this might be different. In a preliminary study, we used Llama-30B (Touvron et al., 2023a) for dialogue generation and annotation. We asked a student assistant from our lab to assess the results. They constantly rated the Llama-30B dialogues lower in terms of naturalness, coherence, engagement, task coverage, i.e., how close is the generated dialogue to the task description, and (turn) length (see Table 11).

Model	Naturalness	Coherence	Engagement	Task Coverage	Length
GPT-3.5-Turbo	4.40	4.92	1.0	4.68	7,12
LLaMA-30B	3.12	3.52	0.8	3.52	3,24

Table 11: Result of our analysis comparing dialogues generated by GPT-3.5-Turbo and Llama-30B. Except for Engagement and Length, all measurments are based on a Likert scale from 1 (lowest rating) to 5 (highest rating).

We suspect that this is rather due to the differences in model size and context window. While GPT-3.5-Turbo has a context window of 4k tokens, Llama-30B has a context window of only 2k tokens. However, regardless of the model used, LLM-generated data oftentimes suffers from various kinds of hallucinations (Zhang et al., 2023; Ji et al., 2023), which makes data curation with humans inevitable. In our data curation study (Section 6), we learned that this is not only much easier for humans, they are also much more efficient in curating annotated dialogues than collecting and annotating them from scratch. For example, collecting and annotating one dialogue takes on average ten minutes and requires two humans. For GPT-3.5-Turbo it is only 90 seconds. Curating an annotated dialogue took on average four minutes and did not require a partner.

G Additional Details and Experiments

In this section, we provide additional information on our experiments, including hyperparameters, input sequences and the results for our continual learning experiments.

Hyperparameters For the experiments with feedback-free dialogues, we trained all models for five epochs, except for Llama 2 (Touvron et al., 2023b), which was trained for ten epochs, since it took already five epochs to adapt the pretrained model to our prompting mechanism. For the experiments with feedback dialogues, we subsequently trained the best performing feedback-free models for ten epochs using the feedback data (ten epochs, since we have seen further improvements after the fifth epoch).

For all experiments, we used a batch size of 32 and a learning rate of $5e-5$ with no warmup steps. As optimizer, we used the implementation of AdamW (Loshchilov and Hutter, 2019) in Py-

Experiment		Generation Accuracy			Task Completion				Quality	
		F1	BLEU	BertScore	Inform	Success	Intent Acc.	Slot Acc.	Toxicity	Q^2
Flan-T5 Feedback-Free	Flan-T5	41.9	16.2	87.4	84.6	81.2	93.8	64.0	0.02	51.2
Feedback	+Generation Error	41.2 (–0.7)	15.4 (–0.8)	87.6 (+0.2)	89.2 (+4.6)	84.1 (+2.9)	95.0 (+1.2)	86.7 (+22.7)	0.02	54.6 (+3.4)
	+User Reaction	42.7 (+0.8)	16.0 (–0.2)	88.9 (+2.5)	83.2 (–1.2)	81.7 (+0.5)	91.8 (–2.0)	87.0 (+23.0)	0.02	53.9 (+2.7)
	+Generation Error +User Reaction	42.9 (+1.0)	16.3 (+0.1)	87.8 (+0.4)	89.5 (+4.9)	86.2 (+5.0)	97.0 (+3.2)	88.5 (+24.5)	0.02	58.2 (+7.0)
GPT-2 Feedback-Free	GPT-2	50.1	27.1	90.5	83.2	82.1	96.5	93.3	0.02	22.1
Feedback	+Generation Error	53.1 (+3.0)	30.2 (+3.1)	90.6 (+0.1)	89.2 (+6.0)	90.1 (+8.0)	98.0 (+1.5)	92.9 (–0.4)	0.02	25.6 (+3.5)
	+User Reaction	50.8 (–0.3)	26.8 (–0.3)	89.8 (–0.7)	88.5 (+2.3)	86.1 (+4.0)	94.6 (–1.9)	93.9 (+0.6)	0.02	29.4 (+7.3)
	+Generation Error +User Reaction	51.5 (+1.4)	26.2 (–0.9)	90.2 (–0.3)	90.2 (+7.0)	87.9 (+5.8)	98.0 (+1.5)	94.9 (+1.6)	0.02	28.1 (+6.0)
Llama 2 Feedback-Free	Llama 2	31.6	7.5	87.1	80.1	84.5	52.3	56.7	0.01	35.1
Feedback	+Generation Error	31.4 (–0.2)	8.3 (+0.8)	87.7 (+0.6)	87.6 (+7.5)	85.4 (+0.9)	59.3 (+7.0)	61.2 (+4.5)	0.01	36.7 (+1.6)
	+User Reaction	32.2 (+0.6)	8.5 (+1.0)	84.6 (–2.5)	86.9 (+6.8)	84.2 (–0.3)	60.1 (+7.8)	62.4 (+5.7)	0.02 (–0.01)	35.9 (+0.8)
	+Generation Error +User Reaction	32.5 (+0.9)	8.5 (+1.0)	84.1 (+3.0)	83.5 (+3.4)	84.1 (–0.4)	60.7 (+8.4)	59.6 (+2.9)	0.02 (–0.01)	36.2 (+1.1)

Table 12: Results on the feedback validation data (averaged over three runs). We use the feedback-free models as deltas for calculating the differences.

torch[16]. Except for Llama 2, we fully-finetuned all models. For Llama 2, we only finetuned the LoRA (Hu et al., 2022) weights, using a rank of 8, an alpha of 16, and a dropout rate of 0.05.

Input Sequences Each model used in this work requires a different input sequence. In general, the components of the input sequence depend on the features used (e.g., user emotions or demographic information). Figure 17 shows the input sequence used for training and inference using Flan-T5 (Chung et al., 2022). Additionally added source data is highlighted in blue in the figures below.

<knowledge> {document} <user_persona> {demographic information} <user_emotion> {emotion} <error_text> {error text} <user_reaction> {user reaction} <dialogue> {context} </s>

Figure 17: Input sequence for Flan-T5.

The target sequence includes the intent, slot values, and system response. It is basically the same as the last part of the input sequence for GPT-2 (Radford et al., 2019), which is shown in Figure 18 (starting from <intent>, but without the special token).

For inference with GPT-2, we use the same input sequence as for Flan-T5 (Figure 17). For Llama-2 (Touvron et al., 2023b), Figure 19 shows the input sequence.

In contrast to Flan-T5 and GPT-2, we use an instruction for training and inference with Llama 2.

[16] AdamW in the Pytorch documentation (last accessed 30 January 2024).

<knowledge> {document} <user_persona> {demographic information} <user_emotion> {emotion} <error_text> {error text} <user_reaction> {user reaction} <dialogue> {context} <intent> {intent} <slots> {slots} <system> {target} <|endoftext|>

Figure 18: Input sequence for GPT-2.

Given is the following task-oriented document-grounded dialogue (<dialogue>) between a human user (<user>) and a virtual agent (<system>). Previously, this conversation went wrong because the virtual agent made a statement that was contextually incorrect ({error text}). The human user reacted accordingly ({user reaction}). Generate the user's intent (<intent>), extract the slot values (<slots>) and generate the next system utterance by considering the user's emotion ({emotion}), persona ({demographic information}) and the following document: {document} <dialogue> {context} <intent> {intent} <slots> {slots} <system> {target}

Figure 19: Input sequence for Llama 2.

For inference, we only use the sequence up to the dialogue context (similar to GPT-2).

Validation Results Table 12 shows the results of the feedback-trained models on the feedback validation data. We use the models from Table 4, but apply them to the validation data of the feedback dialogues. In contrast to the results achieved on the human-human test dialogues, the feedback-trained models outperform the baseline models in terms of generation accuracy in most cases.

Continual Learning Experiments Table 13 shows the results of our continual learning experiments using the most promising configurations from Section 7 and the human-human test dialogues. For each model, we use the best performing feedback-free model from Section 7 (Table 4) as

a starting point. We train the models sequentially with each version of the dialogues, starting with Version 2 and once with annotations for implicit user feedback (Feedback) and once without (No Feedback). The rest of the training procedure and hyperparameter configuration corresponds to what is described above. Due to the large number of experiments, we only present single run results here (the results in Section 7 were averaged over three runs).

Interestingly, the results are rather mixed. We observe a tendency for the task completion metrics to improve with each version of the dialogues, especially when using the annotations for implicit user feedback. The same applies to factual consistency (Q^2 (Honovich et al., 2021)). Overall, the positive impact is not as large as in our experiments in Table 4 (Section 7), but from our perspective this does not mean that continual learning from such data is not beneficial. As discussed in Section 2, there are works available that show the opposite. We rather attribute this to the varying annotation quality between the different versions of the dialogues in FEDI (see Section 5), which we are planning to address in future work.

Model	Experiment	Generation Accuracy			Task Completion				Quality	
		F1	BLEU	BertScore	Inform	Success	Intent Acc.	Slot Acc.	Toxicity	Q^2
Version 2										
No Feedback	Flan-T5 +Emotions	52.8	29.4	89.4	86.5	83.2	86.8	85.0	0.02	55.6
	GPT-2 +Emotions +Demographics	35.4	9.9	85.0	86.4	83.9	89.0	81.6	0.02	31.7
	Llama 2 +Emotions	45.7	25.1	85.4	88.4	86.1	40.6	39.8	0.02	29.5
Feedback	Flan-T5 +Emotions +Generation Error +User Reaction	54.9	33.0	89.7	95.6	93.2	87.5	85.3	0.02	59.8
	GPT-2 +Emotions +Demographics +Generation Error +User Reaction	35.4	10.3	85.3	84.7	83.3	93.0	85.0	0.02	28.9
	Llama 2 +Emotions +Generation Error	40.8	19.6	84.9	91.1	94.9	51.2	52.6	0.01	30.3
Version 3										
No Feedback	Flan-T5 +Emotions	52.5 (–0.3)	31.5 (+2.1)	88.8 (–0.6)	86.9 (+0.4)	85.4 (+2.2)	80.8 (–6.0)	85.0	0.02	55.3 (–0.3)
	GPT-2 +Emotions +Demographics	33.7 (–1.7)	9.6 (–0.3)	84.3 (–0.7)	86.5 (+0.1)	83.3 (–0.6)	89.0	83.4 (+1.8)	0.02	29.2 (–2.5)
	Llama 2 +Emotions	30.0 (–15.7)	15.3 (–9.8)	83.0 (–2.4)	87.4 (–1.0)	85.2 (–0.9)	38.5 (–2.1)	37.6 (–2.2)	0.02	30.4 (+0.9)
Feedback	Flan-T5 +Emotions +Generation Error +User Reaction	49.2 (–5.7)	29.8 (–3.2)	88.3 (–1.4)	96.1 (+0.5)	95.1 (+1.9)	82.3 (–5.2)	84.6 (–0.7)	0.02	58.8 (–1.0)
	GPT-2 +Emotions +Demographics +Generation Error +User Reaction	36.1 (+0.7)	12.0 (+1.7)	85.1 (–0.2)	94.7 (+10.0)	89.1 (+5.8)	93.0	85.0	0.02	33.2 (+4.3)
	Llama 2 +Emotions +Generation Error	39.4 (–1.4)	21.2 (+1.6)	74.9 (–10.0)	92.0 (+0.9)	90.6 (–4.3)	55.1 (+3.9)	58.6 (+6.0)	0.01	32.4 (+2.1)
Version 4										
No Feedback	Flan-T5 +Emotions	49.6 (–3.2)	28.7 (–0.7)	88.3 (–1.1)	85.9 (–0.6)	83.2	81.0 (–5.8)	82.9 (–2.1)	0.02	57.3 (+1.6)
	GPT-2 +Emotions +Demographics	33.4 (–2.0)	10.2 (+0.3)	84.8 (–0.1)	87.1 (+0.7)	83.6 (–0.3)	86.0 (–3.0)	84.6 (+3.0)	0.02	31.4 (–0.3)
	Llama 2 +Emotions	28.7 (–17.0)	14.5 (–10.6)	85.4	90.1 (+1.7)	86.7 (+0.6)	41.0 (+0.4)	42.3 (+2.5)	0.02	31.6 (–2.1)
Feedback	Flan-T5 +Emotions +Generation Error +User Reaction	50.6 (–4.3)	32.7 (–0.3)	88.6 (–2.1)	98.1 (+2.5)	96.2 (–3.0)	81.3 (–6.2)	85.0 (–0.3)	0.02	60.5 (+0.7)
	GPT-2 +Emotions +Demographics +Generation Error +User Reaction	34.9 (–0.5)	11.7 (+1.4)	87.5 (+2.2)	99.3 (+14.6)	97.5 (+14.2)	91.0 (–2.0)	85.5 (+0.5)	0.02	34.9 (+6.0)
	Llama 2 +Emotions +Generation Error	40.1 (–0.7)	15.4 (–4.2)	82.1 (–2.8)	94.5 (+3.4)	96.1 (+1.2)	54.4 (+3.2)	60.2 (+7.6)	0.01	33.9 (+3.6)

Table 13: Results achieved on the test data for each stage. We use the respective models from Version 2 as deltas for calculating the difference in Version 3 and 4.