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

Anonymous ACL submission

Abstract

Implicit user feedback, user emotions and demographic information have shown to be promising sources for improving the accuracy and user engagement of responses generated by dialogue systems. However, the influence of such information on task completion and factual consistency, which are important criteria for task-oriented and document-grounded dialogues, is not yet known. To address this, we introduce FEDI, the first English task-oriented document-grounded dialogue dataset annotated with this information. Our experiments with Flan-T5, GPT-2 and Llama 2 show a particularly positive impact on task completion and factual consistency. Participants in our human evaluation reported that the responses generated by the feedback-trained models were more informative (Flan-T5 and GPT-2), more relevant and more factual consistent (Llama 2).[1]

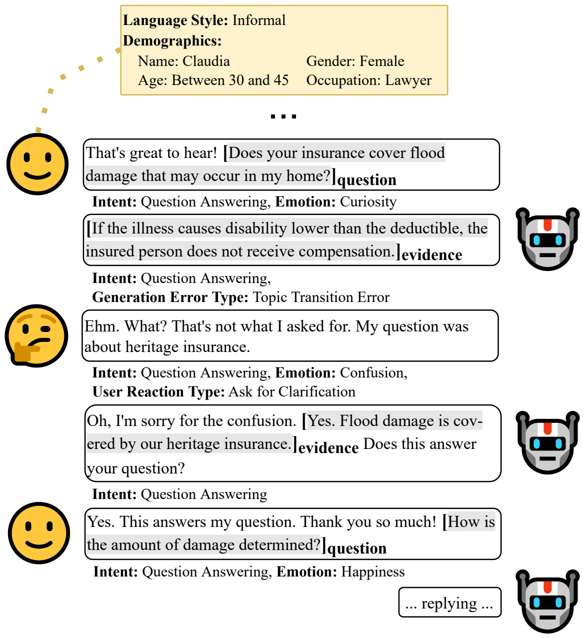


Figure 1: A feedback dialogue from FEDI. User emotion and implicit user feedback annotations (generation error and user feedback types) are beneath the utterances.

1 Introduction

Implicit user feedback (Xu et al., 2023b; Veron et al., 2021; Hancock et al., 2019), such as clarification questions, user emotions (Hwang et al., 2023; Rashkin et al., 2019; Hsu et al., 2018) and demographic information (Lee et al., 2022; Zhang et al., 2018), such as age or language style, are promising sources for improving the accuracy and user engagement of responses generated by dialogue systems. For example, in the second utterance of Figure 1, the system generates a response unrelated to the user's question, which affects her emotional state. She asks the system for clarification, getting a more satisfying response. This makes her happy and she continues the conversation. However, we do not know to what extent the generated response contributes to achieving the user's goal and reflects the underlying knowledge source. This is commonly referred to as task completion and factual consistency. Both are important criteria for task-oriented and document-grounded dialogue systems (Nekvinda and Dušek, 2021; Honovich et al., 2021; Budzianowski et al., 2018), but the impact of implicit user feedback, user emotions and demographic information on them is an open research question.

To address this gap, we introduce the FEDI dataset. Following recent research that includes information-seeking in task-oriented dialogues (Taranukhin et al., 2024; Braunschweiler et al., 2023; Feng, 2021; Campos et al., 2020), e.g., for handling multi-domain scenarios, FEDI provides annotations for required knowledge documents and is the first English task-oriented document-grounded dialogue dataset annotated with implicit user **F**eedback, **E**motions and **D**emographic **I**nformation. FEDI allows us to in-

[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 to other synthetic datasets generated by large language models (LLMs) in terms of avg. turn and utterance length. It also has a higher lexical diversity than many of the crowdsourced datasets[2].

vestigate the impact of this information on task completion and factual consistency of responses generated by dialogue systems, for which we use Flan-T5 (Chung et al., 2022), GPT-2 (Radford et al., 2019) and Llama 2 (Touvron et al., 2023b) in this work. We use GPT-3.5-Turbo[3] to generate and annotate the training and validation data for FEDI. We 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, user emotions and demographic information. It is comparable to other related datasets in terms of size, lexical diversity and dialogue length (see Table 1).

2. New experimental insights, e.g., on the positive impact of feedback data on task completion and factual consistency, and on how humans perceive the responses generated by the resulting models.

3. A framework for generating and annotating task-oriented document-grounded feedback-annotated dialogue data. Our analysis provides insights into the quality of the generated annotations.

[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.

2 Related Work

Learning from emotions has a positive impact on generation accuracy and user engagement in dialogue systems (Firdaus et al., 2020; Rashkin et al., 2019; Hsu et al., 2018). This also applies to demographic information (Hwang et al., 2023; Lee et al., 2022; Siddique et al., 2022; Luo et al., 2019; Zhang et al., 2018) and implicit user feedback, although the latter requires the (continual) training of a model with feedback data that must first be collected in human interaction (Xu et al., 2023a,b; Ung et al., 2022; Veron et al., 2021; Mazumder et al., 2020; Wang et al., 2019; Hancock et al., 2019). Table 1 compares the datasets from the aforementioned works, if publicly available. For task-oriented dialogues, EmoWOZ (Feng et al., 2022) provides annotations for user emotions, but focuses only on emotion recognition. The datasets annotated with implicit user feedback (Xu et al., 2023a,b; Ung et al., 2022) do not distinguish different types of implicit user feedback, as suggested by Higashinaka et al. (2021) and Petrak et al. (2023), and are tailored to their use case. Besides, most of the available datasets are the result of costly crowdsourcing efforts, often leading to datasets of varying quality, e.g., due to methodical artifacts or annotator biases (Parmar et al., 2023; Yang et al., 2023; Thorn Jakobsen et al., 2022; Prabhakaran et al., 2021). Recent works suggest synthetic data generation using large language models, especially GPT-3.5-Turbo, as a more efficient approach to generate high-quality dialogue data (Kim et al., 2023;

Li et al., 2023a,b; Lee et al., 2022). However, they also show that such models are heavily dependent on detailed instructions and task descriptions for this purpose and still have a tendency to generate biased, hallucinated or harmful output (Yang et al., 2023; Ji et al., 2023; Zhang et al., 2023; Malaviya et al., 2023).

We introduce FEDI, a task-oriented document-grounded dialogue dataset, which gets the best of both worlds. We use GPT-3.5-Turbo to generate and annotate training and validation data and recruit human annotators for quality assessment, curation, and to collect a separate set of test dialogues. To cover a broad variety of types for implicit user feedback and generation errors, we use the taxonomies proposed by Petrak et al. (2023).

3 FEDI

FEDI covers four use cases from three customer service domains, including postal, receptionist and insurance services. For postal 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 insurance services, 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 knowledge documents. Appendix A describes the tasks in more detail, including slots, intents and document sources.

Implicit User Feedback (GE, F) We use the taxonomies proposed by Petrak et al. (2023) to generate and annotate generation errors (GE) and subsequent implicit user feedback (F). They distinguish ten types of generation errors. Nine of which are relevant for FEDI, such as Attribute Error, Factually Incorrect or Lack of Sociality. For implicit user feedback, they distinguish five types, e.g., Ask for Clarification, Ignore and Continue or Repeat or Rephrase. Definitions, further details and examples can be found in Appendix B.

Demographic Information (DI) 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 (E) 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 found in related work (Kim et al., 2023; Rashkin et al., 2019) which we assume to be relevant for the tasks represented in FEDI, 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 $C = [T_0, ..., T_{t-1}]$, 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 gender, age range, occupation, name, and language style of the user. 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 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. For each step, we require GPT-3.5-Turbo to return the results in a predefined JSON scheme. If in one step the generation does not match this

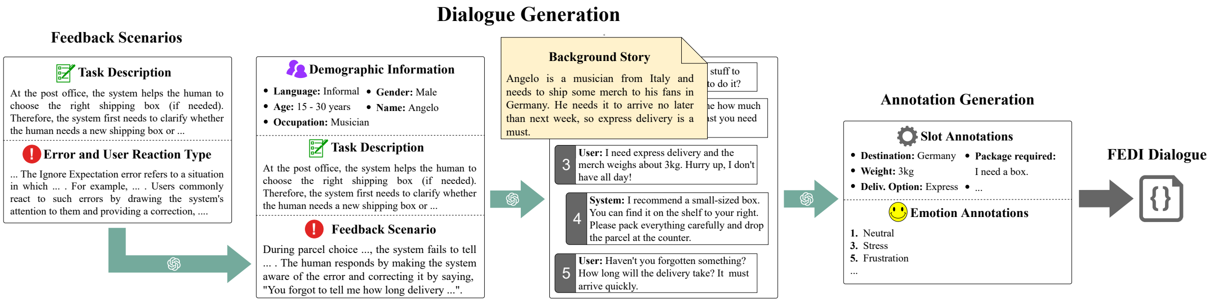


Figure 2: Overview of our framework for generating and annotating dialogues. (the green arrow) symbolizes GPT-3.5-Turbo. The generation of feedback dialogues requires feedback scenarios as additional source. For question answering dialogues, we include the respective documents in the task description.

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

The procedure for dialogue generation is basically the same for feedback-free and feedback dialogues. It starts in the second box from the left in Figure 2. We provide GPT-3.5-Turbo with randomly sampled demographic information for the user and a task description, which also includes the role of the starting actor, i.e., user or system. Feedback dialogues require feedback scenarios as additional source (Section 4.2). A task description describes the flow of events and information which needs to be conveyed by each role to fulfill the task. For question answering, this 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 demographic information to generate a background story to guide the conversation, such as depicted in the center of the figure. We require the model to return the utterance-level annotations for intents (not included in the figure) 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]. To prevent hallucinations, we instruct the model to only copy values from the dialogue and to return the annotations on utterance-level. For emotion annotations, we instruct the model to predict the emotion for each user utterance, given the dialogue and our emotion taxonomy.

4.2 Feedback Dialogues

Feedback Scenarios A feedback scenario describes a generation error and the following implicit user feedback. Figure 2 shows an example in the second box from the left. For generation (first box), we provide GPT-3.5-Turbo with the task description and a list of randomly sampled generation error and implicit user feedback 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[5].

Feedback Dialogue Generation For feedback dialogues, 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, e.g., *Yes, I can help you send a parcel to Paris*, a subsequent user utterance that reflects the user reaction, e.g., *No, the destination is London, not Paris!* and a following system utterance that addresses the user reaction, e.g., *Apologies for the mistake. Thank you for correcting me. The destination is London, United Kingdom. Now, please provide me with the weight of the package*. We consider the dialogue as Version 1 and generate three additional versions of the same dialogue, each resolving one of the feedback scenarios.

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

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

Resolving Feedback Scenarios Figure 3 illustrates the idea. For each version, we first mask the affected system utterance and generate a replacement using the preceding dialogue context and task-specific information. Next, we drop the following two utterances, since they are directly related to the generation error. This way, the dialogue remains coherent and the conversation continues with the next regular user utterance[6].

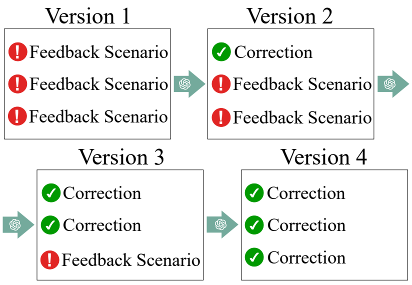


Figure 3: Feedback dialogue generation. Each version solves one of the feedback scenarios from Version 1. See Appendix C (Figure 13) for an example dialogue.

We continue the process until all feedback scenarios 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 FEDI Analysis

FEDI consists of 8,852 dialogues, divided into 1,988 feedback-free dialogues, including 326 for testing, and 6,864 feedback dialogues (1,716 in four versions). The test dialogues were collected human-human by eight computer science students. We provide details on recruitment, salary, procedure, and our experiences and findings from collecting and annotating dialogue data with humans vs. LLMs in Appendix D.

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.

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[7]. For the feedback-free dialogues, we distinguish between generated dialogues (Gen.) and test dialogues (Test).

We observe large differences between (1) question answering and the other tasks, and (2) the generated dialogues and the human-collected test dialogues. We attribute this to the different slot annotation schemes. While question answering has a rather simple slot annotation scheme (see Appendix A), the slots in the other tasks often depend on the background story, e.g., 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 into account and occasionally omit slots that are not required based on the background story, 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 sometimes do not contain all the required slot values. 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[8].

As expected, negative emotions are more common in feedback dialogues. Happiness in feedback dialogues is mostly observed when the system addresses the implicit user feedback. This is similar for curiosity, although we also observe this emotion when the system suddenly changes the topic.

[6]We experimented with different ideas for resolving feedback scenarios (see Appendix C), but the naive approach described here turned out to be the most reliable.

[7]Hallucinated slot values, i.e. slot annotations that do not occur in the respective utterance, were small in number and are counted as missing.

[8]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).

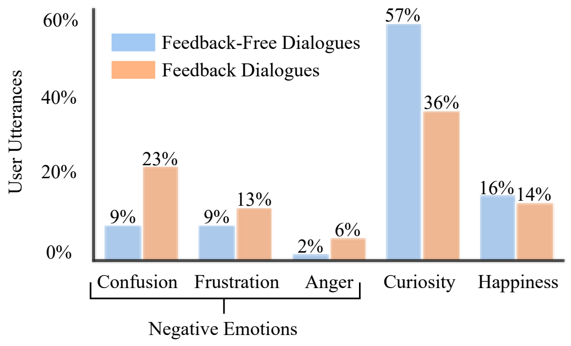


Figure 4: Ratio of the most commonly observed user emotions in FEDI (excluding the Neutral emotion).

While this emotion usually fits the context, it can also be the result of insufficient information in the emotion annotation instruction, as we only use the dialogue context as additional information and no further examples (see Appendix C).

Feedback Scenarios Figure 5 shows the distribution of user reaction types in relation to generation error types represented in the feedback scenarios of the feedback dialogues. It shows that our approach for generating feedback scenarios mostly resulted in meaningful combinations of generation error and user reaction types. For example, Factually Incorrect 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.

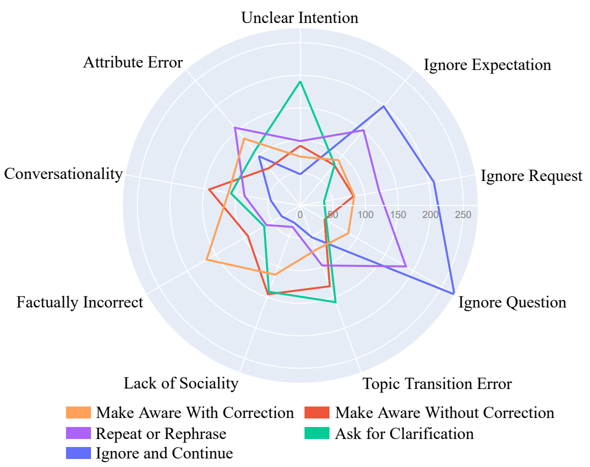


Figure 5: Distribution of user reaction types in relation to generation error types represented in feedback scenarios.

6 Quality Control for FEDI

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 implicit user feedback type annotations in 380 feedback dialogues (see Appendix D.3 for the procedure). We used INCEpTION (Klie et al., 2018) as a platform for this study. We calculate the inter-annotator agreement (IAA) using Krippendorff's Alpha (Krippendorff, 2006) with a nominal weighting function (as provided in the platform). Table 3 shows the results[9].

	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 annotations. We found that most of them are parcel shipping dialogues, which has a comparatively complex annotation scheme (see Appendix A). 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, annotators reported 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. We attribute this to our observation from Section 5 (GPT-3.5-Turbo strictly follows the slot annotation scheme, even if the values are not in the dialogue). Emotions, whose perception is very subjective, are the most frequently changed annotation type (on average 2.09 times per affected dialogue), whereby the originally annotated emotion was often 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

[9] Overall, 26 dialogues were reported as off-topic (13/480 feedback-free and 13/380 feedback). They are not considered in these results.

Experiment		Task Completion				Quality		Generation Accuracy		
		Inform	Success	Intent Acc.	Slot Acc.	Q²	Toxicity	F1	BLEU	BertScore
Flan-T5 Feedback-Free	Flan-T5	86.7	85.9	54.8	60.9	52.7	0.02	45.0	20.0	88.3
	+Emotions	**83.9** **(–2.8)**	**83.2** **(–2.7)**	**61.2** **(+6.4)**	**58.3** **(–2.6)**	**57.5** **(+4.8)**	**0.02**	**46.7** **(+1.7)**	**21.0** **(+1.0)**	**88.9** **(+0.6)**
	+Demographics	87.0 (+0.3)	86.0 (+0.1)	33.5 (–21.3)	29.3 (–31.6)	54.5 (+1.8)	0.03 (+0.01)	43.2 (–1.8)	18.4 (–1.6)	87.7 (–0.6)
	+Emotions +Demographics	85.3 (–1.4)	85.1 (–0.8)	43.9 (–10.9)	36.7 (–24.2)	56.4 (+3.7)	0.02	44.2 (–0.8)	19.1 (–0.9)	88.1 (–0.2)
Feedback	+Generation Error	96.8 (+10.1)	92.7 (+6.8)	72.5 (+17.7)	76.7 (+15.8)	56.9 (+4.2)	0.02	41.4 (–3.6)	19.8 (–0.2)	87.8 (–0.5)
	+User Reaction	96.6 (+9.9)	94.1 (+8.2)	69.0 (+14.2)	76.2 (+15.3)	56.3 (+3.6)	0.02	41.3 (–3.7)	19.3 (–0.7)	87.6 (–0.7)
	+Generation Error +User Reaction	**96.9** **(+10.2)**	**95.3** **(+9.4)**	**83.5** **(+28.7)**	**77.2** **(+16.3)**	**60.2** **(+7.5)**	**0.02**	**44.4** **(–0.6)**	**22.1** **(+2.1)**	**88.2** **(–0.1)**
GPT-2 Feedback-Free	GPT-2	88.3	81.6	78.7	69.6	28.1	0.02	34.9	10.4	87.1
	+Emotions	84.1 (–4.2)	83.8 (+2.2)	75.4 (–3.3)	67.3 (–2.3)	26.7 (–1.4)	0.02	35.1 (+0.2)	10.4	87.1
	+Demographics	80.2 (–8.1)	80.2 (–1.4)	69.3 (–9.4)	57.5 (–12.1)	26.3 (–1.8)	0.02	34.6 (–0.3)	10.4	87.1
	+Emotions +Demographics	**85.1** **(–3.2)**	**84.8** **(+3.2)**	**71.6** **(–7.1)**	**66.7** **(– 2.9)**	**29.2** **(+1.1)**	**0.02**	**36.0** **(+1.1)**	**11.4** **(+1.0)**	**87.3** **(+0.2)**
Feedback	+Generation Error	92.4 (+4.1)	91.7 (+10.1)	84.3 (+5.6)	79.3 (–9.7)	30.9 (+2.8)	0.02	29.2 (–5.7)	8.0 (–2.4)	86.2 (–0.9)
	+User Reaction	98.9 (+10.6)	96.5 (+14.9)	83.0 (+4.3)	80.3 (+10.7)	32.3 (+4.2)	0.02	30.0 (–4.9)	8.3 (–2.1)	86.3 (–0.8)
	+Generation Error +User Reaction	**94.7** **(+6.4)**	**93.3** **(+11.7)**	**88.0** **(+9.3)**	**80.8** **(+11.2)**	**35.5** **(+7.4)**	**0.01** **(–0.01)**	**30.3** **(–4.6)**	**9.7** **(–0.7)**	**86.4** **(–0.7)**
Llama 2 Feedback-Free	Llama 2	85.9	81.2	37.6	39.2	28.3	0.02	29.3	7.1	86.1
	+Emotions	**89.3** **(+3.4)**	**85.3** **(+4.1)**	**40.2** **(+2.6)**	**41.3** **(+2.1)**	**18.7** **(–9.6)**	**0.01** **(–0.01)**	**36.3** **(+7.0)**	**14.9** **(+7.8)**	**85.4** **(–0.7)**
	+Demographics	85.6 (–0.3)	82.5 (+1.3)	37.1 (–0.5)	40.1 (+0.9)	21.3 (–7.0)	0.02	33.8 (+4.5)	4.5 (–2.6)	86.5 (+0.4)
	+Emotions +Demographics	86.7 (+0.8)	87.9 (+6.7)	41.4 (+3.8)	39.6 (+0.4)	20.6 (–7.7)	0.03 (+0.01)	28.8 (–0.5)	5.6 (–1.5)	81.3 (–4.8)
Feedback	+Generation Error	93.1 (+7.2)	95.7 (+14.5)	54.8 (+17.2)	59.6 (+20.4)	29.1 (+0.8)	0.01 (–0.01)	24.1 (–5.2)	7.9 (+0.8)	77.4 (–8.7)
	+User Reaction	**94.9** **(+9.0)**	**93.2** **(+12.0)**	**63.5** **(+25.9)**	**70.1** **(+30.9)**	**27.1** **(–1.2)**	**0.02**	**24.5** **(–4.8)**	**6.9** **(–0.2)**	**78.8** **(–7.3)**
	+Generation Error +User Reaction	82.4 (–3.5)	83.6 (+2.4)	46.3 (+8.7)	47.2 (+8.0)	33.5 (+5.2)	0.03 (+0.01)	25.0 (–4.3)	9.2 (+2.1)	80.1 (–6.0)

Table 4: The results of our experiments. We use the pretrained models finetuned on the feedback-free dialogues as deltas. The best-performing models are highlighted. Learning from user emotions (+Emotions) positively impacts the generation accuracy. The demographic information (+Demographics) is of minor importance. The feedback experiments show that 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 (Q^2).

2 (Touvron et al., 2023b) (7B, plain pretrained version)[10]. We first finetune the pretrained models using the feedback-free dialogues and include the user emotions, demographic information and documents as part of the input sequences. For Llama 2, we only finetune the LoRA (Hu et al., 2022) weights in our experiments. We use the best performing feedback-free models for experiments with the feedback dialogues. We provide additional details in the Appendix, including hyperparameters (F.1), input sequences (F.2) and experiments for continual learning from feedback data (H).

Evaluation Metrics We use Inform and Success (Budzianowski et al., 2018) to measure Task Completion. Additionally, we measure the correctness of the predicted intent and slot values (intent and slot accuracy). We use Q^2 (Honovich et al., 2021) to measure the factual consistency of the generated responses (in question answering). Since the generation errors in FEDI include social aspects (see Appendix B), we use Perspective API to measure their toxicity, and F1-Score, BLEU(-n) (Papineni et al., 2002) and BertScore (Zhang et al., 2020)[11] to measure their generation accuracy.

Results Table 4 shows the results achieved in the test dialogues (averaged over three runs). The feedback-free experiments show that including user emotions has the most positive impact. It improves the generation accuracy and factual consistency for Flan-T5 (Chung et al., 2022) and GPT-2 (Radford et al., 2019) (here in combination with demographic information), and the generation accuracy and task completion for Llama 2 (Touvron et al., 2023b). The feedback experiments show improved task completion and factual consistency across all models. We assume that the generation errors and user reactions used in training served as negative examples, helping the models to learn to generate more accurate intents and slots and responses that

[10] 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).

[11] For Inform and Success, we use the implementation from Nekvinda and Dušek (2021) as a reference. For Q^2, we use the reference implementation which is available in GitHub. Perspective API is a free-to-use service provided by Google and Jigsaw. We measure the F1-Score based on the overlapping tokens in target and prediction. For BLEU (Papineni et al., 2002) and BertScore (Zhang et al., 2020), we use the implementation from the HuggingFace evaluation library v0.4.1 and with $n = 4$ for BLEU (last access to all resources on 04 January 2024).

better reflect the knowledge documents (see Appendix F.3 for examples). An analysis on dialogue type level (Appendix F.4) also shows an increased generation accuracy for Flan-T5 and GPT-2, but only for question answering. We assume this is due to the knowledge document, which is part of the context in the input sequence and helps generating responses close to the target sequences. Related works that report increased generation accuracy usually use similar mechanisms to regulate the impact of feedback training (Xu et al., 2023a,b; Ung et al., 2022). We found that the responses generated for the other tasks, in which we do not use knowledge documents, still fit the context well but often deviate from the target sequences. For Flan-T5 and GPT-2, this is also reflected in the behavior of the F1-Score, which measures word overlapping and is more affected than BLEU (Papineni et al., 2002) and BertScore (Zhang et al., 2020). For Llama 2, we found that the length of the generated responses sometimes deviates from that of the target sequences in both the feedback-free and feedback experiments. The predicted intent and slot values also sometimes deviate from the target values. Therefore, we assume this as the reason for the performance deviations from GPT-2. Regarding toxicity, we did not observe any negative impact from including generation errors, except for some outliers in Flan-T5 and Llama 2 (see appendix F.5).

Human Evaluation We conduct a human evaluation to investigate how humans perceive the impact of feedback training. We recruited 42 participants from Prolific[12] and asked them to rate the human likeness, relevancy, sociality, engagement, and factual consistency of the responses generated for 300 randomly sampled test dialogues in the feedback and feedback-free experiments highlighted in Table 4 (50 test dialogues from each experiment). We used a Likert scale from one to five for each attribute (with one as the lowest value). We received 40 valid submissions (we checked them manually in detail). Thus, each dialogue was rated by at least five participants. We provide more details on our rating scheme, annotator background and procedure in Appendix G. Table 5 shows the results.

For Flan-T5 (Chung et al., 2022) and GPT-2 (Radford et al., 2019), annotators reported that the responses generated by the feedback models are more informative (which is not captured by the scores), but do not always cover the knowledge document as well as the responses from the feedback-free models (Flan-T5). They also reported them to contain more counter-questions, which is actually desirable, but is often perceived as disruptive and sometimes inattentive, and more direct, which is sometimes perceived as unfriendly. This is reflected in the slightly lower scores for relevancy and sociality (see Appendix G.2 for examples). For Llama 2 (Touvron et al., 2023b), annotators reported some responses of the feedback-free model as illogical, unrelated to the dialogue context and factually incorrect. The responses generated by the feedback model were rated much higher, especially their relevancy and factual consistency. The IAA is rather low for most measures, which we attribute to their subjectivity and the diversity of annotators.

Model	Human Likeness		Relevancy		Sociality		Engagement		Factual Consistency	
	Rating	IAA	Rating	IAA	Rating	IAA	Rating	IAA	Rating	IAA
Flan-T5										
Feed.-Free	3.41	0.15	4.12	0.19	4.66	0.15	3.56	0.15	4.12	0.50
Feedback	**3.27**	0.09	**3.99**	0.19	4.56	0.10	3.57	0.11	4.02	0.50
GPT-2										
Feed.-Free	3.25	0.12	3.97	0.19	4.70	0.14	3.60	0.19	3.63	0.26
Feedback	**4.02**	0.12	3.88	0.25	**4.58**	0.14	3.52	0.12	3.64	0.41
Llama 2										
Feed.-Free	3.0	0.19	3.31	0.28	4.49	0.11	3.16	0.25	2.74	0.40
Feedback	**3.12**	0.12	**3.87**	0.29	**4.64**	0.12	**3.54**	0.23	**3.69**	0.37

Table 5: Results of our human evaluation. If statistically significant, they are printed in bold. (independent two-sample t-test, $p \leq 0.05$). We calculated IAA again using Krippendorff's Alpha (Krippendorff, 2006)[13].

8 Conclusion

We introduce FEDI, the first English task-oriented document-grounded dialogue dataset annotated with implicit user feedback, user emotions and demographic information. Our analysis shows the usefulness of our framework for generating feedback-annotated dialogues and that FEDI is comparable to other related datasets. Our experiments show that learning from implicit user feedback improves task completion and factual consistency. Humans perceive the responses generated by feedback models as more informative (Flan-T5 and GPT-2), more relevant and more factually consistent (Llama 2). However, our results also show room for improvements in future work on learning from feedback data, e.g., the varying impact on generation accuracy in different dialogue tasks and the influence on the tone of generated responses.

[12] Prolific is a widely used crowdsourcing platform for scientific research (last accessed 08 May 2024).

[13] We used SciPy v1.13.0 for the t-test (last accessed 08 April 2024). For Krippendorff's Alpha, we used K-Alpha Calculator (Marzi et al., 2024) (interval weighting).

9 Limitations

Taxonomies Used The taxonomies used for generating implicit user feedback, user emotions and demographic information only reflect subsets of possible values. They are not exhaustive. For example, we do not consider country or race for demographic information, or other emotions than those that seemed meaningful to us in the context of this work.

Synthetically Generated Data 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). Model-specific bias could also be a factor, which we haven't investigated further. Although our analysis shows that the generated annotations are of high quality and we have invested a lot of effort in developing the instructions used, some values may be incorrect or inappropriate in context, e.g., in the case of user emotions. This also applies to the slot and intent annotations, where analysis has shown that human annotators can react more flexibly to the dialogue background. In contrast, GPT-3.5-Turbo focuses completely on the instruction and tends to return placeholder values in case of doubt. In addition, some of these dialogues may seem artificial and unnatural due to potentially conflicting demographic information, e.g., language style contradicting 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. Although we conducted a fairly extensive human curation study in which we did not observe these issues, a more thorough review of the whole dataset would be required for a final assessment.

To solve feedback scenarios, we experimented with different ideas to incorporate the feedback into regenerating the affected system utterance. However, this led to unnatural and inconsistent dialogues, which is why we decided to use the naive approach described in the paper. As a result, the regenerated system utterances may not always directly reflect the feedback.

Toxicity Through Learning From Generation Errors In our feedback experiments, we also use generation errors for learning. Since they also include social aspects, such as disrespectful or toxic response behavior, we used Perspective API to analyze the toxicity in generated responses. Although conspicuous responses were very rare, 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 offensive to specific demographic groups, which are not marked by the detector (Cheng et al., 2023).

Human Evaluation We conducted the human evaluation as a crowdsourcing study and recruited 42 participants so that each dialogue was evaluated seven times. Some participants submitted their assessment far below the time limit, which is why we carefully checked each individual submission. Due to deviations from our rating scheme, we had to discard two submissions, which is why 100 of the 300 dialogues considered received fewer than seven ratings. Another limitation is the study design. We only considered the quality of the generated responses and not that of the generated slot and intent values. During the study, we found that our rating scheme has limitations as well. For example, hallucinations were not considered as a separate measure. Some annotators reported them as comments to the affected dialogues. However, the number was very small and we did not notice any additional cases when checking the submissions.

References

Norbert Braunschweiler, Rama Doddipatla, Simon Keizer, and Svetlana Stoyanchev. 2023. Evaluating large language models for document-grounded response generation in information-seeking dialogues. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 46–55, Prague, Czech Republic. Association for Computational Linguistics.

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.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

7302–7314, Online. Association for Computational Linguistics.

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.

Song Feng. 2021. DialDoc 2021 shared task: Goal-oriented document-grounded dialogue modeling. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 1–7, Online. Association for Computational Linguistics.

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.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. 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. Emotion-Lines: 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. 2023a. Norm-Dial: 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.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, 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.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

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.

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.

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.

Maksym Taranukhin, Sahithya Ravi, Gábor Lukács, Evangelos Milios, and Vered Shwartz. 2024. Empowering air travelers: A chatbot for canadian air passenger rights. *CoRR*, abs/2403.12678.

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. SaFeRDialogues: 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. 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.

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 topping up a 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.

Receptionist Services For receptionist services, FEDI only includes one task: access control. Table 7 shows the slots for this task. 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 and 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[14]. Table 8 lists the slots. In the past, customers called their insurance agent or visited their local bank branch for all questions related to such topics. Today, it is more common to talk to chatbots or other service

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

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.

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.

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 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[15], 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[16]. 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 generation errors and implicit user feedback, we use the error and user reaction type taxonomies proposed by Petrak et al. (2023). For generation errors in system utterances they define the following nine error types as relevant for task-oriented and 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.

[15]Available in GitHub (last accessed on 16 January 2024).
[16]Published by babymed.com (last accessed 12 February 2024).

- **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 rare in task-oriented and 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

Prompt engineering played a major role in this work. The instructions used to generate the dialogues and annotations were continuously improved in an iterative process to generate valid data within the given parameters. This section only focuses on the final instructions used in this work. Additionally 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.

Feedback Dialogues Figure 8 shows the instruction for the generation of feedback scenarios, which are required as an additional source for feedback dialogues.

{list of error type names} are common generation errors in dialogues.
{list of error type definitions}
Users commonly react to such errors by {list of 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 error type.

Figure 8: Instruction for generating the feedback scenarios.

For each feedback dialogue, 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.

Generate an erroneous long and in-depth dialogue (at least 13 turns) 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.

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.

Resolving Feedback Scenarios For each feedback dialogue (Version 1), we generate three additional versions of the same dialogue, each resolving one of the feedback scenarios. For this, we experimented with different ideas:

- 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.

They all resulted in inconsistent dialogues and off-topic or unnatural system utterances. We found that using the dialogue context up to the affected system utterance, masking and regenerating this utterance (in a friendly and polite manner), leads to the best matching and most coherent replacements. Figure 10 shows the instruction.

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 for regenerating the system utterance to replace the one with the generation error.

It includes the dialogue context, the name of the task and the document if the task is question answering. Although GPT-3.5-Turbo has a long context length, we found that including the full task descriptions was distracting rather than improving the replacements. This means that the model can only use internal knowledge and information from the dialogue context for generating the replacements, which sometimes had a negative impact on the completeness of the slot annotations, e.g., for parcel shipping and topping up a prepaid SIM card (see Section 5).

After replacing the affected system utterance, we regenerate its slot values. We remove the following two utterances to ensure the dialogue flow is not corrupted (since they directly refer to the generation error). The conversation then continues with the next regular user utterance. Figure 13 shows an example dialogue from FEDI to illustrate this procedure.

Slot Annotations Figure 11 shows our instruction for generating the slot annotations.

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.

Given is the following dialogue between a dialogue system and a person (user):
{dialogue}
The dialogue 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.

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.

D Test Data Collection and Curation Study

We hired student assistants for our test data collection and curation study. In this section, we want to provide more insights into the application criteria, hiring procedure, and data collection.

D.1 Application Criteria and Hiring Procedure

To participate, 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 advertisement to our university. Also, we did not consider gender. We asked all applicants who fulfilled those criteria to participate in a recruitment test, in which

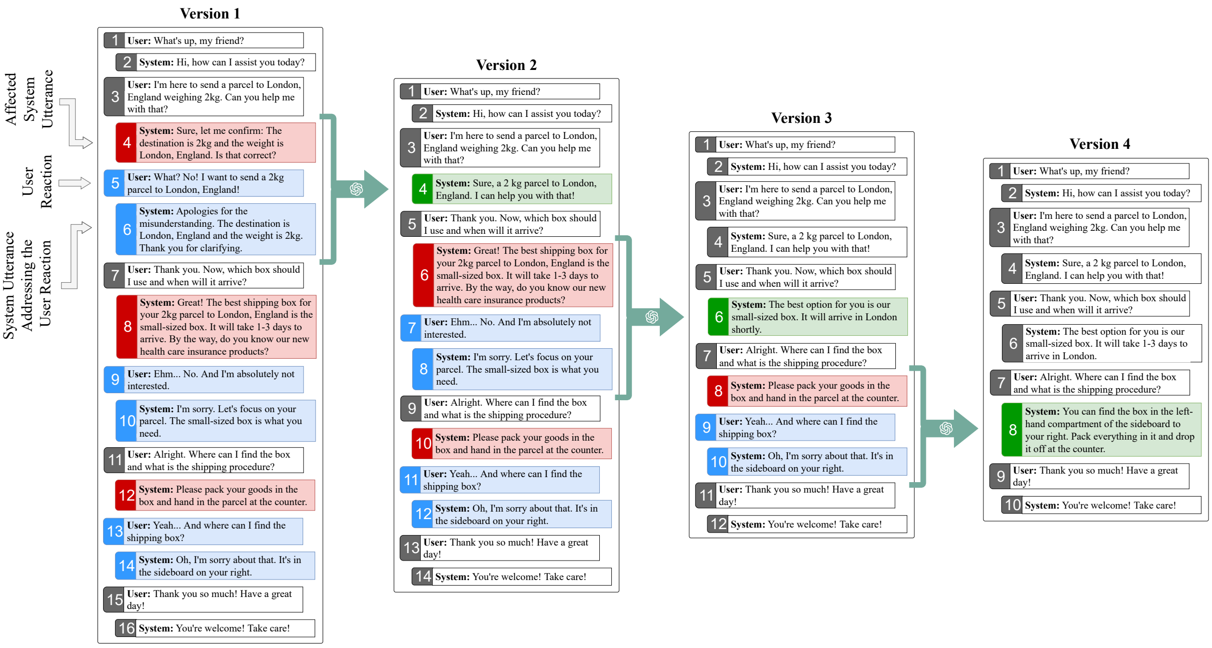


Figure 13: Example dialogue from FEDI for illustrating our approach for resolving feedback scenarios. In each version, we keep the previous part of the dialogue, regenerate the affected system utterance and drop the following two utterances (the user reaction and the system utterance which addresses the user reaction), since they are directly related to the generation error.

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 completeness, (3) number of turns per dialogue, (4) avg. utterance length.

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

D.2 Test Data Collection

The test data for FEDI was collected by eight computer science students in overall 136 paid working hours. 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 used a self-developed web-based platform that allows to collect and annotate dialogues between two humans. Figure 14 shows the user interface.

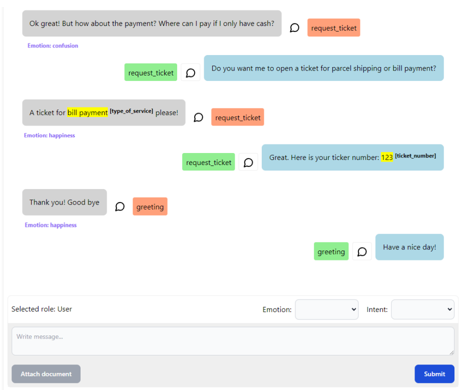


Figure 14: The user interface of the data collection platform used to collect the test data.

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. User emotion annotations are colored purple. For Question Answering, the chatpane also allows attaching a document to a message (a text file).

D.3 Data Curation Study

For the curation of the generated data, the procedure was different for feedback-free and feedback dialogues. For feedback-free dialogues, we asked the annotators to assess and correct (add/modify/delete) the generated slot and intent annotations per utterance, and their completeness on dialogue level (with respect to the task description). We assigned the annotators to the tasks and asked them to work through the corresponding dialogues provided in INCEpTION (Klie et al., 2018). Figure 15 shows the user interface for intent and slot annotation curation.

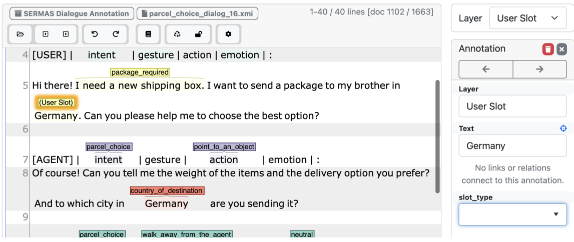


Figure 15: User interface for intent and slot annotation curation in INCEpTION. It's a parcel shipping dialogue and the annotation for country of destination (*Germany* in line eight) is misplaced, because this slot should be provided by the user, who has already mentioned it in line five.

For feedback dialogues, we asked the annotators to assess and correct the annotations for implemented feedback scenarios, i.e., the annotation for error type in the affected system utterance and the user reaction type in the subsequent user utterance. In addition to the information available in the user interface, we provided the annotators with the task descriptions (Appendix A). For feedback dialogues, we also provided them with the definitions of error and user reaction types (Appendix B).

D.4 Dialogue Collection: Human vs. LLM

In our human test 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 9).

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 9: 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.

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 10 shows the distribution of dialogues in the dataset. Test refers to the human-human collected test data.

Demographic Information Figure 16 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 10: Data splits included in FEDI and their sizes.

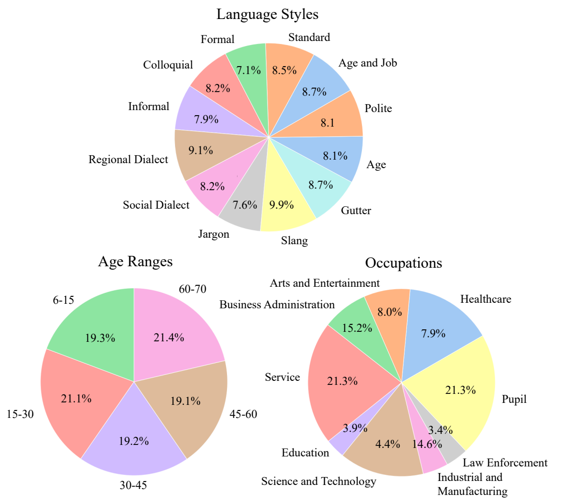


Figure 16: The distribution of persona attributes represented in the background stories (excluding human-human test dialogues).

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[17]. 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 in the entire dataset.

Emotions The chart in Figure 17 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. Other refers to emotions that are represented $\leq$ 5%, including Anger, Disgust, Fear, Surprise, and Stress.

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

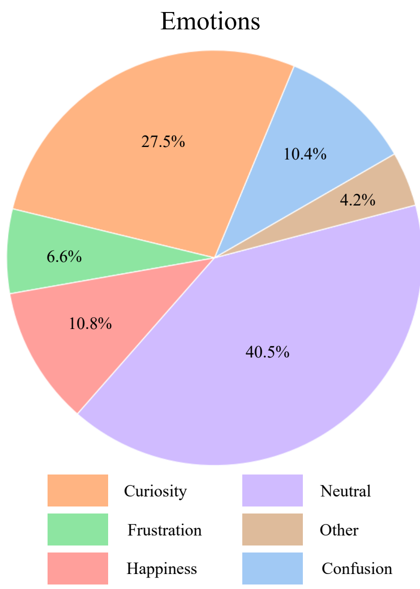


Figure 17: Illustration of the distribution of emotions in FEDI.

Feedback Scenarios Overall, we generated 4,714 feedback scenarios that are included in the feedback dialogues of Version 1. Figure 18 shows the distribution of generation error and user reaction types.

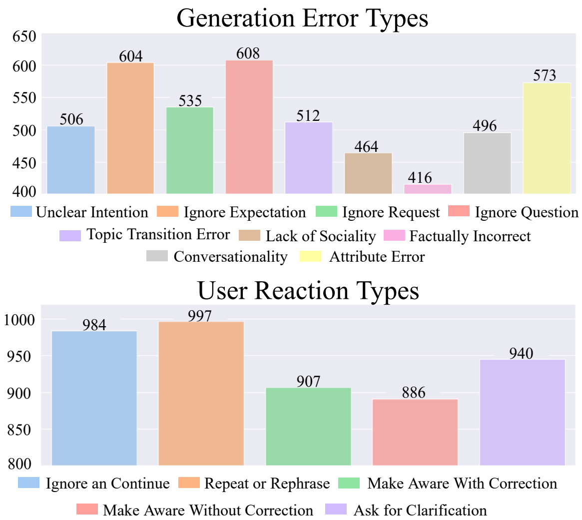


Figure 18: Distribution of generation error and user reaction types in the feedback dialogues of FEDI.

Given that most of the dialogues are about question answering (Table 10), it is not surprising that Ignore Question is the most frequent error type. Table 11 shows the ten most commonly observed error and user reaction type combinations.

Ignore Question and Ignore Request are two of the most frequent error types. While we observe the first one more common in question answering

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

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 Experimental Details and Further Results Analysis Experiments

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

F.1 Hyperparameter

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 (we used the plain pretrained model in our experiments, not the one finetuned on dialogue data). 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 Pytorch[18]. 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.

F.2 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 19 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 19: 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 20 (starting from <intent>).

<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 20: Input sequence for GPT-2.

For inference with GPT-2, we used the same sequence as for Flan-T5. For Llama 2 (Touvron et al., 2023b), Figure 21 shows the sequence.

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 21: Input sequence for Llama 2.

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

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

F.3 Feedback Data as Negative Samples

We attribute the performance improvements in the feedback experiments to the additional context provided by the generation error and user reaction. We assume they serve as a negative example during training and help the models to learn to generate more accurate intents, slots and responses that better reflect the knowledge documents. This section provides an example from our experiments to support this intuition.

The dialogue is a question answering dialogue from the financial domain, and the subject of the conversation is using the Postepay Evolution Card. This is the respective knowledge document: *The Postepay Evolution Card allows the Cardholder, within the limits of the amount available, to make cash withdrawals and payment transactions (with the exception of purchases by mail or telephone and all transactions that are not authorized online, i.e., at the same time as the payment) in Italy and abroad. In addition, the card allows for the payment of highway tolls on the enabled sections.* In the dialogue, the user wants to know if they can use the card for online purchases. The target intent is *question answering* (the respective task) and the target slot is the user's question (*I also wanted to inquire about the usability of the Postepay Evolution Card for online purchases*). The feedback scenario is the following: The system provided the user with a factually incorrect response (*Yes, you can also use the card for purchases by mail or telephone.*), and the user responded with a correction (*Sorry, but that's not true. The Postepay Evolution Card cannot be used for purchases by mail or telephone.*). We generated the next system utterance using the configuration and model from the feedback-free and feedback experiments highlighted in Table 4.

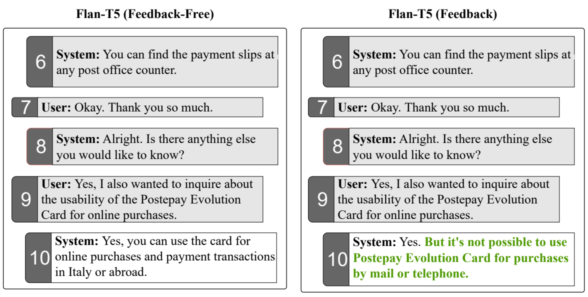


Figure 22: Example for Flan-T5. The feedback-free model was trained using user emotions as an additional input signal. The feedback model was then additionally trained using generation errors and user reactions.

Figure 22 shows the results for Flan-T5 (Chung et al., 2022). The responses are focused on the question and do not provide any additional details, but the feedback model focuses more on the information from the knowledge document. The feedback-free model predicted Bill Form Payment Procedure as the slot, which is incorrect, and set the complete user utterance as the value. Question would have been the correct slot type and the target value does not include the complete user utterance, but only the part after the comma. The feedback model predicted both correctly.

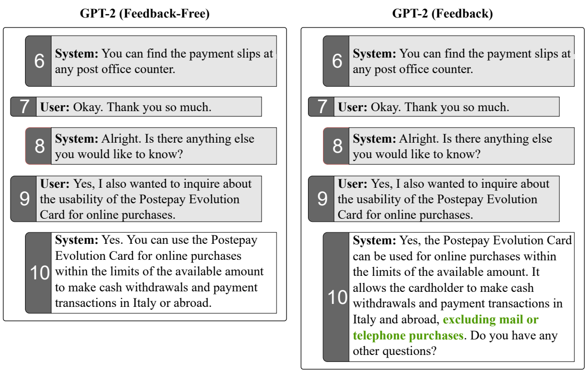


Figure 23: Example for GPT-2. The feedback-free model was trained using user emotions and demographic information as additional input signals. The feedback model was then additionally trained using generation errors and user reactions (like Flan-T5).

Figure 23 shows the results for GPT-2 (Radford et al., 2019). The responses provide more details and read more naturally. However, the response from the feedback model is closer to the knowledge document and ends with a call for interaction. Both the feedback-free and feedback models correctly predicted the intent, slot values, and types.

Figure 24 shows the responses from Llama 2 (Touvron et al., 2023b). In terms of content, they hardly differ from the system utterances generated by GPT-2, they are just not as concise. However, in contrast to Flan-T5 and GPT-2, Llama 2 mispredicted the intent and slot values in both cases. The feedback model predicted information retrieval as value for intent. For the slot value, it did not extract the user's question from their utterance (although this instruction is included in the input sequence, see Appendix F.2). Instead, the model returned a reformulation: *Can I use the Postepay Evolution Card to make purchases by mail or telephone?*

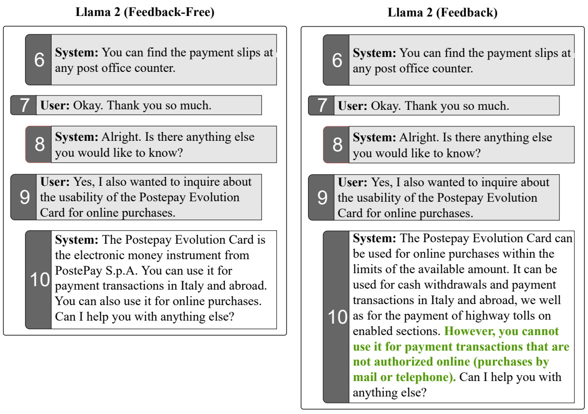


Figure 24: Example for Llama 2. The feedback-free model was trained using user emotions as additional input signals. The feedback model was then additionally trained using just user reactions.

F.4 Generation Accuracy Performance Gap

Table 12, Table 13 and Table 14 show the results from Table 4 divided into question answering (QA) and the other tasks (Others), including parcel shipping, topping up a prepaid SIM card, and access control. For feedback, we only consider the best configuration for each model. As in Table 4, we use the respective base models as deltas and highlight the best performing configurations. Since the FEDI test split contains only 119 ToD dialogs and 207 QA dialogs (see Table 10), we randomly selected 119 samples from question answering to ensure comparability.

	Experiment	F1	BLEU	BertScore
QA	Flan-T5	47.6	25.9	88.1
	+ Emotions	**53.5** (+5.9)	**31.3** (+5.4)	**89.7** (+1.6)
	+ Demographics	52.2 (+4.6)	30.2 (+4.3)	88.9 (+0.8)
	+ Emotions + Demographics	50.0 (+2.4)	28.2 (+2.3)	88.7 (+0.6)
	+ Emotions + Generation Error + User Reaction	48.8 (+1.2)	32.0 (+6.1)	89.2 (+1.1)
Others	Flan-T5	33.6	5.2	87.2
	+ Emotions	**36.7** (+3.1)	**6.8** (+1.6)	**88.4** (+1.2)
	+ Demographics	32.9 (–0.7)	5.6 (+0.4)	86.5 (–0.7)
	+ Emotions + Demographics	32.3 (–1.3)	5.8 (+0.6)	87.6 (+0.4)
	+ Emotions + Generation Error + User Reaction	30.9 (–2.7)	5.4 (+0.2)	85.5 (–1.7)

Table 12: Generation accuracy in the question answering and task-oriented dialogues for Flan-T5.

For Flan-T5 (Chung et al., 2022) and GPT-2 (Radford et al., 2019), the results in question answering are usually much higher than for the other tasks. A manual analysis revealed that the responses generated for question answering are primarily summaries of the corresponding knowledge documents, like the target sequences for this task. Therefore, we assume that this is the reason for the comparatively good generation accuracy for this dialogue type. We also assume that these knowledge documents serve as a regulating mechanism when learning from feedback, similar to those used in related work (Xu et al., 2023b,a; Ung et al., 2022) (see also the examples in Appendix F.3). We found that the responses generated for the other tasks in these experiments still fit the context well, but often deviate from the target sequences. This is also expressed in the behavior of the scores. While the F1-Score measures word overlap and is therefore more affected, the other metrics, which focus more on contextual similarity, are less affected. We assume that the results could be different if we could find a similar guiding mechanism (or guiding source) for the other tasks. The task descriptions from dialogue generation (Appendix A) could be an interesting starting point for such experiments, as they provide a pattern for the expected dialogue flow and information about the required slot values.

	Experiment	F1	BLEU	BertScore
QA	GPT-2	35.3	11.3	86.4
	+ Emotions	**40.6** (+5.3)	**16.2** (+4.9)	**89.6** (+3.2)
	+ Demographics	38.6 (+3.3)	10.1 (–1.2)	89.6 (+3.2)
	+ Emotions + Demographics	40.9 (+5.6)	11.2 (–0.1)	89.2 (+2.8)
	+ Emotions + Demographics + Generation Error + User Reaction	37.4 (+2.1)	12.1 (+0.8)	89.0 (+2.6)
Others	GPT-2	34.4	11.1	86.3
	+ Emotions	34.1 (–0.3)	10.7 (–0.4)	86.1 (–0.2)
	+ Demographics	31.7 (–2.7)	9.8 (–1.3)	85.9 (–0.4)
	+ Emotions + Demographics	**34.5** (+0.1)	**11.6** (+0.5)	**86.4** (+0.1)
	+ Emotions + Demographics + Generation Error + User Reaction	26.2 (–8.2)	7.9 (–3.2)	85.6 (–0.7)

Table 13: Generation accuracy in the question answering and task-oriented dialogues for GPT-2.

For Llama 2, we do not observe any significant change in performance for either question answering or the other tasks. We attribute this to the observation made in Section 7 that the system utterances generated by Llama 2 usually significantly deviate in length from the target sequence (although we used the same number of new tokens in all our experiments), resulting in lower word-overlapping scores.

	Experiment	F1	BLEU	BertScore
QA	**Llama 2**	**34.0**	**12.7**	**84.8**
	+ Emotions	32.1 (−1.9)	10.6 (−2.1)	84.9 (+0.1)
	+ Demographics	31.5 (−2.5)	6.2 (−6.5)	86.1 (+1.3)
	+ Emotions + Demographics	29.1 (−4.9)	6.1 (−6.6)	85.9 (+1.1)
	+ Emotions + User Reaction	24.4 (−9.6)	7.9 (−4.8)	76.0 (−8.8)
Others	Llama 2	26.5	8.0	86.6
	+ Emotions	28.3 (+1.8)	5.9 (−2.1)	85.4 (−1.2)
	+ Demographics	28.9 (+2.4)	5.6 (−2.4)	85.7 (−0.9)
	+ Emotions + Demographics	**27.4 (+0.9)**	**8.3 (+0.3)**	**86.3 (−0.3)**
	+ Emotions + User Reaction	22.9 (−3.6)	4.7 (−3.3)	87.4 (+0.8)

Table 14: Generation accuracy in the question answering and task-oriented dialogues for Llama 2.

F.5 Impact of Learning From Generation Errors on Toxicity

Although their share is small (Lack of Sociality in Figure 18), the generation errors in FEDI contain potentially toxic and disrespectful language. Table 4 shows that the toxicity of generated responses is generally negligible (values are ≤ 0.03). However, we observe some outliers in the Flan-T5 (Chung et al., 2022) and Llama 2 (Touvron et al., 2023b) feedback models which score ≥ 0.1. For example, Flan-T5 + Emotions + Generation Error + User Reaction once generated *Alright, that's a start. What else? And don't forget, I need it in simple terms. None of that fancy shit.* to request for missing information in the case of parcel shipping. For Llama 2 + Emotions + Generation Error, toxicity scores ≥ 0.1 are sometimes observed in the case of question answering, e.g., *The Legal Protection does not apply to events resulting from popular riots, acts of terrorism, vandalism, earthquakes, strikes and lock-outs, possession or use of radioactive substances, disputes concerning family, inheritance and gift law, tax and administrative disputes, events resulting from popular riots, insurrections, military operations, acts of terrorism, vandalism, earthquakes, strikes and lock-outs.* However, we consider these false positives, as they may contain critical terms but do not offend the user personally. Overall, generated system utterances with a toxicity score ≥ 0.1 are extremely rare ($\leq 0.1\%$ of the responses generated with these models in the test data).

In the feedback-free experiments, we did not observe any generated system utterance with a toxicity score ≥ 0.1.

G Crowdsourcing Study

We did a crowdsourcing study to investigate how humans perceive the impact of feedback training. For this, we hired 42 crowdworkers on Prolific for an hourly salary of 9,00$ (the hourly salary recommended by the platform). Our requirement for participation was as follows:

- Fluent in English.
- At least 10 previous submissions to other studies on Prolific.
- Approval rate of at least 90%.

We did not restrict participation to US citizens. We also did not consider gender, age or other educational background. We had no further influence on the allocation of participants. To manage this (and the payment) is the purpose of Prolific. The participants were forwarded to Google Forms, which we used to implement our study (see appendix G.1).

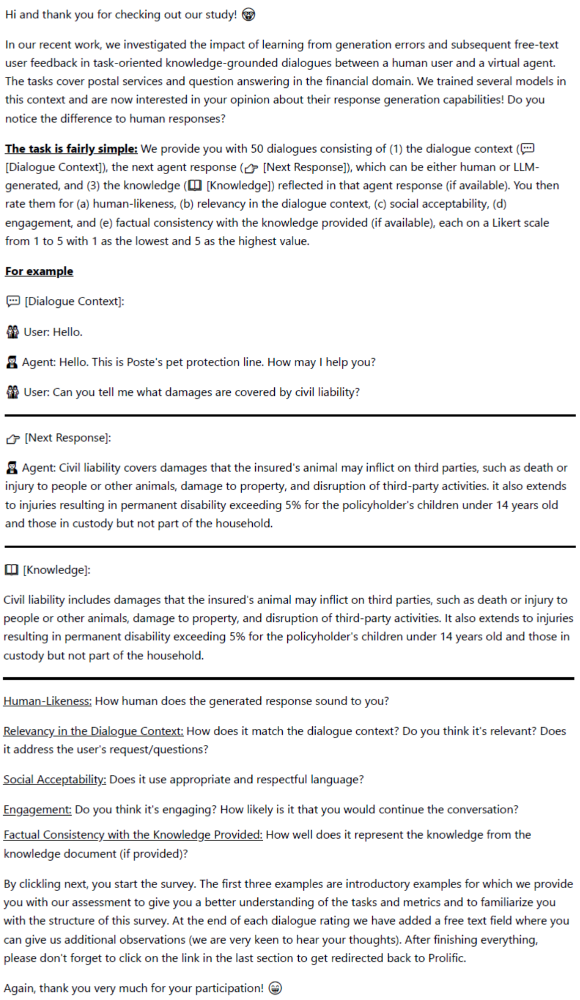
Overall, from the 42 people who decided to participate, 23 were from South Africa and 19 from european countries. 24 of the participants were female. 18 were male. The average age was 28.54 years. The youngest person was 21 years old. The oldest person 62. We did not conduct any recruitment test in advance. Instead, we provided the participants with three test samples in the live study so that they could become familiar with the task and our rating scheme. We reviewed all submissions in detail and decided to exclude the results of two participants, as they contained predominantly incomprehensible ratings (we paid them nevertheless).

G.1 Implementation and Procedure

We implemented the crowdsourcing study using Google Forms[19], using one section per dialogue. At the beginning of the survey, we provided the participants with extensive instructions describing the task and the rating scheme (see Figure 25). Figure 26 shows an example dialogue from our study.

For each dialogue, we presented the annotators the dialogue context, generated response and knowledge document (in the case of question answering), but did not indicate whether the response was generated by a human or language model. We used Python scripts and the Google Forms API

[19] Google Forms is a survey management software that is part of the free, web-based Google Docs Editor Suite from Google (last accessed 09 May 2024).


Hi and thank you for checking out our study!

In our recent work, we investigated the impact of learning from generation errors and subsequent free-text user feedback in task-oriented knowledge-grounded dialogues between a human user and a virtual agent. The tasks cover postal services and question answering in the financial domain. We trained several models in this context and are now interested in your opinion about their response generation capabilities! Do you notice the difference to human responses?

The task is fairly simple: We provide you with 50 dialogues consisting of (1) the dialogue context ([Dialogue Context]), the next agent response ([Next Response]), which can be either human or LLM-generated, and (3) the knowledge ([Knowledge]) reflected in that agent response (if available). You then rate them for (a) human-likeness, (b) relevancy in the dialogue context, (c) social acceptability, (d) engagement, and (e) factual consistency with the knowledge provided (if available), each on a Likert scale from 1 to 5 with 1 as the lowest and 5 as the highest value.

For example

[Dialogue Context]:

User: Hello.

Agent: Hello. This is Poste's pet protection line. How may I help you?

User: Can you tell me what damages are covered by civil liability?

[Next Response]:

Agent: Civil liability covers damages that the insured's animal may inflict on third parties, such as death or injury to people or other animals, damage to property, and disruption of third-party activities. it also extends to injuries resulting in permanent disability exceeding 5% for the policyholder's children under 14 years old and those in custody but not part of the household.

[Knowledge]:

Civil liability includes damages that the insured's animal may inflict on third parties, such as death or injury to people or other animals, damage to property, and disruption of third-party activities. It also extends to injuries resulting in permanent disability exceeding 5% for the policyholder's children under 14 years old and those in custody but not part of the household.

Human-Likeness: How human does the generated response sound to you?

Relevancy in the Dialogue Context: How does it match the dialogue context? Do you think it's relevant? Does it address the user's request/questions?

Social Acceptability: Does it use appropriate and respectful language?

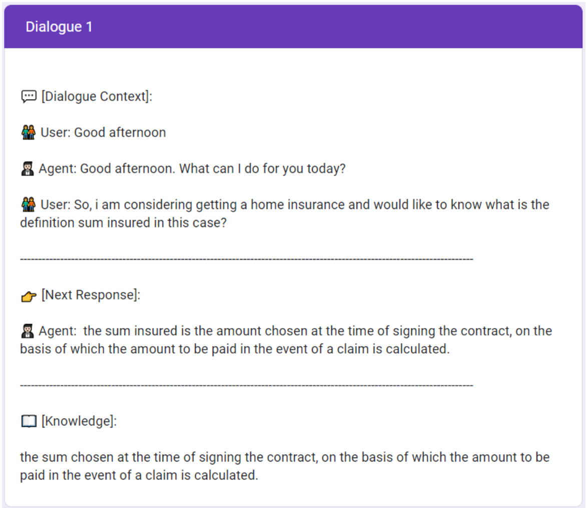
Engagement: Do you think it's engaging? How likely is it that you would continue the conversation?

Factual Consistency with the Knowledge Provided: How well does it represent the knowledge from the knowledge document (if provided)?

By clickling next, you start the survey. The first three examples are introductory examples for which we provide you with our assessment to give you a better understanding of the tasks and metrics and to familiarize you with the structure of this survey. At the end of each dialogue rating we have added a free text field where you can give us additional observations (we are very keen to hear your thoughts). After finishing everything, please don't forget to click on the link in the last section to get redirected back to Prolific.

Again, thank you very much for your participation!

Figure 25: Task Description for our crowdsourcing study.


Dialogue 1

[Dialogue Context]:

User: Good afternoon

Agent: Good afternoon. What can I do for you today?

User: So, i am considering getting a home insurance and would like to know what is the definition sum insured in this case?

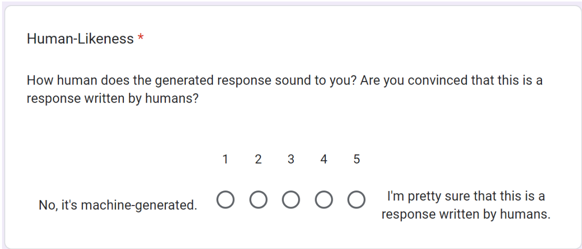
[Next Response]:

Agent: the sum insured is the amount chosen at the time of signing the contract, on the basis of which the amount to be paid in the event of a claim is calculated.

[Knowledge]:

the sum chosen at the time of signing the contract, on the basis of which the amount to be paid in the event of a claim is calculated.

Figure 26: Example dialogue from our crowdsourcing study. Each dialogue was represented as a separate section in a Google Forms survey.


Human-Likeness *

How human does the generated response sound to you? Are you convinced that this is a response written by humans?

1 2 3 4 5

No, it's machine-generated. ○ ○ ○ ○ ○ I'm pretty sure that this is a response written by humans.

Figure 27: Linear scale for human-likeness from our crowdsourcing study.

to automatically create and fill surveys with 50 dialogues randomly sampled from the 300 pre-selected test dialogues. Below the dialogues, we added the rating forms using linear scales from one to five. Figure 27 shows an example.

We asked the annotators to rate the generated responses ([Next Response]) for the following attributes: human likeness (how human does the generated response sound?), relevancy in the dialogue context (does it match the dialogue context? does it address the user's concern?), sociality (does it use appropriate and respectful language?), factual consistency (how well does it represent the knowledge from the document?), and engagement (do you think it is engaging? would you like to continue the conversation?). Filling the rating forms was mandatory. At the end of each section we added a free text field in which they were asked to provide us with additional observations (if any).

G.2 Examples

In this section, we provide examples to illustrate the observations reported by the annotators in our crowdsourcing study. The responses generated by the models used are highlighted in green in the figures. Figure 28 shows the context of a dialogue and the response that was generated by the Flan-T5 (Chung et al., 2022) feedback model. While the annotators agreed that the information presented in the response is correct, they reported in their comments that they felt it was not inviting to continue the conversation.

It answers the question, but does not contain any further request for interaction. Figure 29 shows a response generated by the GPT-2 (Radford et al., 2019) feedback model. This is one of the responses reported as less attentive. The user asks for information about insurance for home damages and focuses on houses in Italy in utterance five. The model does not pick up this information and returns a counter-question asking the user whether the house is in Italy, the Republic of San Marino or

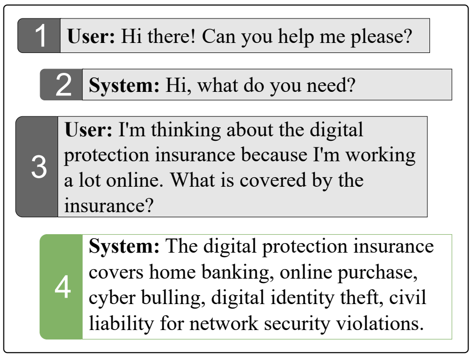


Figure 28: A response generated by the Flan-T5 feedback model presented to the annotators in our crowdsourcing study.

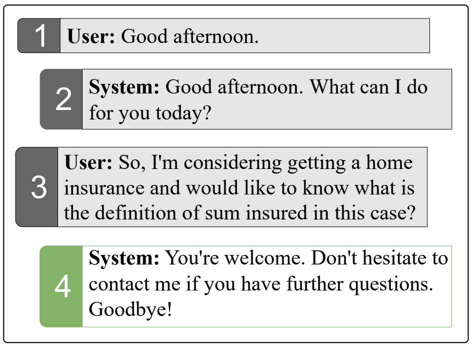


Figure 30: A response generated by the Llama 2 feedback-free model presented to the annotators in our crowdsourcing study.

the Vatican City.

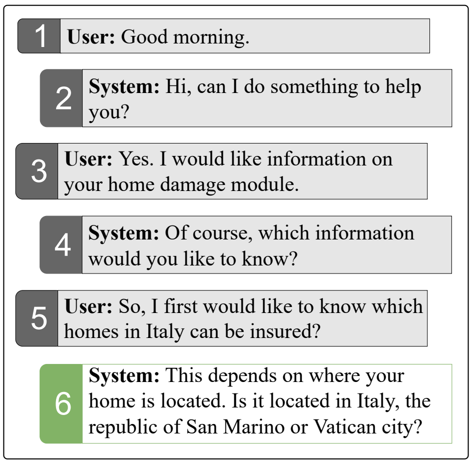


Figure 29: A response generated by the GPT-2 feedback model presented to the annotators in our crowdsourcing study.

Figure 30 shows a sample from the Llama 2 (Touvron et al., 2023b) feedback-free model, which illustrates why annotators reported many of them as illogical or unrelated to the dialogue context. The user is asking about the definition of sum insured in the case of home insurance. Instead of responding to this, the model says goodbye to the user.

We selected these samples because they are exemplary for the observations made by the annotators. The same phenomena were also observed in responses generated to longer dialogue contexts.

H Continual Learning From Feedback Data

Table 15 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 feedback 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 in Appendix F.1. 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)).

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

Table 15: 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.