

Learning From Free-Text Human Feedback – Collect New Datasets Or Extend Existing Ones?

Dominic Petrak¹, Nafise Sadat Moosavi², Ye Tian³, Nikolai Rozanov³, Iryna Gurevych¹

¹UKP Lab, Department of Computer Science, Technical University of Darmstadt, Germany

²Department of Computer Science, The University of Sheffield, United Kingdom

³Wluper, London, United Kingdom

www.ukp.tu-darmstadt.de

Abstract

Learning from free-text human feedback is essential for dialog systems, but annotated data is scarce and usually covers only a small fraction of error types known in conversational AI. Instead of collecting and annotating new datasets from scratch, recent advances in synthetic dialog generation could be used to augment existing dialog datasets with the necessary annotations. However, to assess the feasibility of such an effort, it is important to know the types and frequency of free-text human feedback included in these datasets. In this work, we investigate this question for a variety of commonly used dialog datasets, including MultiWoZ, SGD, BABI, PersonaChat, Wizards-of-Wikipedia, and the human-bot split of the Self-Feeding Chatbot. Using our observations, we derive new taxonomies for the annotation of free-text human feedback in dialogs and investigate the impact of including such data in response generation for three SOTA language generation models, including GPT-2, LLAMA, and Flan-T5. Our findings provide new insights into the composition of the datasets examined, including error types, user response types, and the relations between them¹.

1 Introduction

If a dialog system generates a dissatisfying or erroneous response, e.g., a response with factually incorrect information, users tend to provide a textual description of what went wrong or what they would have expected (See and Manning, 2021; Xu et al., 2023; Ung et al., 2022). This textual description is usually referred to as **free-text human feedback**, and it is an important source to improve dialog systems, to keep them engaging and socially acceptable (Shuster et al., 2022; Christiano et al., 2017; Ouyang et al., 2022; Xu et al., 2023). Learning from such data requires annotations for errors

in system utterances and subsequent free-text human feedback, but available datasets are scarce and typically cover only a small subset of the error types known in conversational AI, such as in the cases of SaFeRDialogues (Ung et al., 2022) and FITS (Xu et al., 2023). As a result, data collection is usually a first step for research on learning from free-text human feedback (Hancock et al., 2019; Veron et al., 2021; Park et al., 2021). To avoid this in future research, recent advances in synthetic dialog generation (Kim et al., 2022; Zheng et al., 2022) could be used to augment existing dialog datasets with annotations for errors and free-text human feedback. However, to assess the feasibility of such an effort, it is important to know the types and frequency of such data included in these datasets.

In this work, we investigate this question for a variety of dialog datasets, including MultiWoZ (Budzianowski et al., 2018), PersonaChat (Zhang et al., 2018), Wizards-of-Wikipedia (Dinan et al., 2019), SGD (Rastogi et al., 2020), BABI (Bordes et al., 2017), and the human-bot split from the Self-Feeding Chatbot (Hancock et al., 2019). For this, we follow a two-step approach in which we first use the Integrated Error Taxonomy proposed by Higashinaka et al. (2021) to study the types of errors in system utterances and subsequent user responses in subsets of these datasets. We use the insights gained during this process to derive a new user response type taxonomy and a modified Integrated Error Taxonomy for the annotation of free-text human feedback in dialog data. In the second step, we use Sentence-Transformer (Reimers and Gurevych, 2019) to identify similar situations in the remaining dialogs of the datasets. For a subsequent statistical analysis, we manually annotate subsets of the identified dialogs with error and user response types. We also use this manually annotated data to investigate the impact of including this data in re-

¹Code and data are available on [GitHub](https://github.com).

sponse generation, using three SOTA language generation models, i.e., GPT-2 (Radford et al., 2019), LLAMA (Touvron et al., 2023), and Flan-T5 (Chung et al., 2022).

We find that types and frequency of errors and user responses that include free-text human feedback largely depend on (1) whether the dialogs are human-human or human-bot, and (2) whether the dialogs are task-oriented, open-domain, or knowledge-grounded. Our analysis and experiments show that using our taxonomies is beneficial for identifying free-text human feedback, and that including such data has a positive impact in response generation.

2 Related Work

Datasets Annotated with Free-Text Human Feedback As of today, dialog datasets with annotations for free-text human feedback are few and mostly focused on specific error types. For example, FITS (Xu et al., 2023), a dataset of 14,000 human-bot dialogs, provides free-text human feedback for response quality and search query errors. SaFeRDialogues (Ung et al., 2022) consists of 7,000 human-bot dialogs, but only addresses toxicity in system utterances. The dataset published with the Self-Feeding Chatbot (Hancock et al., 2019) provides 60,000 human-bot dialogs, partly annotated with response alternatives for dissatisfying system utterances. Other works, such as Park et al. (2021) and Veron et al. (2021) collected free-text human feedback for evaluating their approaches, but never made the data publicly available.

In this work, we investigate the types of errors and subsequent user responses in six widely used dialog datasets, such as MultiWOZ (Budzianowski et al., 2018), PersonaChat (Zhang et al., 2018), and SGD (Rastogi et al., 2020) to assess their extendibility with annotations for learning from free-text human feedback as an alternative to data collection from scratch.

Error and User Response Type Taxonomies

The error taxonomies used to collect the datasets discussed above are very specific which limits their applicability, e.g., in SaFeRDialogues (Ung et al., 2022), they only focus on toxicity. However, errors in conversational AI have long been the subject of research and more comprehensive error taxonomies are already available. For example, Dybkjaer et al. (1996) and Möller et al. (2007) focus on errors in task-oriented dialog systems and distinguish multi-

ple error types with focus on content-related errors and practical aspects. More recently published error taxonomies, such as the Integrated Error Taxonomy proposed by Higashinaka et al. (2021), distinguish levels of errors, which makes them more broadly applicable. The Integrated Error Taxonomy covers 17 error types in four different levels, including utterance, response, context, and society. Regarding user responses subsequent to errors in system utterances, See and Manning (2021) proposed a taxonomy that distinguishes dissatisfaction and unclear user utterances.

In this work, we use the Integrated Error Taxonomy by Higashinaka et al. (2021) to study the errors in the system utterances of 1,200 dialogs from six dialog datasets, based on which we determine (1) the different types of user responses to errors in system utterances, and (2) the limitations of the Integrated Error Taxonomy, such as missing error types.

3 Datasets Examined

Table 1 gives an overview of the datasets examined in this work. Overall, we consider six datasets with dialogs of various types, including task-oriented, open-domain, and knowledge-grounded dialogs, as well as human-human and human-bot dialogs.

Dataset	Type	Mode	# Dialogs
MultiWoZ	Task-Oriented	Human-Human	8,483
SGD	Task-Oriented	Human-Human	16,000
BABI	Task-Oriented	Human-Bot	6,235
PersonaChat	Open-Domain	Human-Human	10,907
Self-Feeding Chatbot	Open-Domain	Human-Bot	60,000
Wizards-of-Wikipedia	Knowledge-Grounded	Human-Human	22,311

Table 1: Overview of the datasets examined in this work.

For task-oriented dialog datasets, we consider MultiWoZ (Budzianowski et al., 2018) (MWOZ), SGD (Rastogi et al., 2020), and BABI (Bordes et al., 2017). They mainly differ in the number of domains included in the dialogs. MWOZ includes seven different domains, SGD 16, and BABI only one (but with dialogs of increasing difficulty). In contrast to MWOZ and SGD, BABI consists of human-bot dialogs. For open-domain dialogs, we consider PersonaChat (Zhang et al., 2018) (PC) and the human-bot split of the Self-Feeding Chatbot (Hancock et al., 2019) (SFC). While PC consists of dialogs between two people who are trying to get to know each other, SFC

consists of human-bot open-domain dialogs². For knowledge-grounded dialogs, we focus on Wizards-of-Wikipedia (Dinan et al., 2019) (**WoW**), which consists of human-human dialogs.

For simplicity, we do not distinguish between human or bot in the following. We always refer to the utterance of the partner as a system utterance.

4 Manual Error Type Analysis and Taxonomies

We first study the errors in system utterances in a randomly sampled set of 1, 200 dialogs (200 dialogs from each of the six datasets), using the Integrated Error Taxonomy proposed by Higashinaka et al. (2021). The taxonomy consists of 17 error types (I1-I17) across four levels: utterance, response, context, and society. They further categorize error types into content violation, i.e., if the error may cause a dialog breakdown, and form violation, i.e., if the system utterance is not interpretable due to massive grammatical problems. Table 2 presents a summary of the error types (see Appendix A for more details).

Level	Form Violation	Content Violation
Utterance	Uninterpretable (I1)	Semantic Error (I3)
	Grammatical Error (I2)	Wrong Information (I4)
Response	Ignore Question (I5)	Ignore Expectation (I9)
	Ignore Request (I6)	
	Ignore Proposal (I7)	
	Ignore Greeting (I8)	
Context	Unclear Intention (I10)	Self-Contradiction (I13)
	Topic transition error (I11)	Contradiction (I14)
	Lack of Information (I12)	Repetition (I15)
Society	Lack of Sociality (I16)	Lack of Common Sense (I17)

Table 2: Error Types included in the Integrated Error Taxonomy.

If we find an error, we analyze the following user response for an **error-indicating phrase**, a short text fragment of arbitrary length that directly addresses the error in the previous system utterance (e.g., *how do you mean*) or indicates user dissatisfaction (e.g., *I don't like that*), and add the surrounding sentence to our list of unique **error-indicating sentences**. We use these error-indicating sentences in Section 5 to explore the remaining dialogs from each dataset for user responses (and thus errors) that are similar to the ones observed in this step³.

²SFC is also partially annotated with alternative responses, but we only consider the non-annotated dialogs in this work.

³We also used the error-indicating phrases instead of the error-indicating-sentences, but found that they are not expressive enough due to their small length (see also Section 4).

Overall, we found 79 errors in system utterances and collected a set of 67 error-indicating sentences with an average sentence length of approximately 6.52 words (see Appendix B for all collected phrases and sentences). Each sentence contains a unique error-indicating phrase with an average length of 3.52 words. Contractions (two words that have been connected, e.g., *don't* or *it's*) are considered as one word. Table 3 shows the distribution of error-indicating sentences across datasets.

Dataset	Task-Oriented			Open-Domain		Know.-Grounded
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
#Sentences	7	0	5	9	36	10

Table 3: Distribution of error-indicating sentences across datasets. *HH* denotes human-human dialogs and *HB* denotes human-bot dialogs.

We find most error-indicating sentences in open-domain and knowledge-grounded datasets, especially in SFC (Hancock et al., 2019).

4.1 Modified Integrated Error Taxonomy

During this study, we found that the Integrated Error Taxonomy (Higashinaka et al., 2021) has weaknesses. Some error types are never observed and others are missing. Based on these insights, we modify the taxonomy for the classification of errors in system utterances. Table 4 shows the result.

Level	Error Type	Description
Response	Ignore Question (E1)	The system utterance ignores the user's question.
	Ignore Request (E2)	The system utterance ignores the user's request to do something.
	Ignore Expectation (E3)	The system utterance does not fulfill the user's expectation.
	Attribute Error (E4)	The system utterance suggests that the system did not get the attributes/slots right.
	Factually Incorrect (E5)	The system utterance contains information that is factually incorrect.
Context	Topic Transition Error (E6)	The system utterance transitions to another / a previous topic without reasonable explanation.
	Conversationality (E7)	The system utterance indicates that the system lost track, e.g., it repeats previous responses (without asking for missing information) or contradicts itself.
	Unclear Intention (E8)	The system utterance suggests that the user's intent was not successfully conveyed.
Society	Lack of Sociality (E9)	The system utterance lacks consideration of social standards, e.g., greetings, is toxic or disrespectful.
	Lack of Common Sense (E10)	The information in the system utterance opposes the opinion of the majority.

Table 4: Modified Integrated Error Taxonomy.

We ignore *Lack of Information* (I12 in Table 2), since it is rarely observed by Higashinaka et al. (2021) and we never observed it in our study. For the same reason, we ignore I1-I3. However, we also found them to be rather ambiguous. For example, the *Semantic Error* (I3) is intended to be used for invalid predicate/argument combinations, such

as situations where a missing letter results in a different meaning (*raining* instead of *training*). This is similar to the *Lack of Common Sense* error type (I17, now E10), since the model is supposed to be aware of the concept, but not in the given context. For *Wrong Information* (I4), we introduce a new error type, *Factually Incorrect* (E5), that extends the original definition for also taking factually incorrect knowledge into account. Furthermore, we ignore *Contradiction* (I14), since it is covered by *Lack of Common Sense* and *Factually Incorrect* errors. We merge *Ignore Proposal* (I7) and *Ignore Request* (I6) into one error type (E2), since both are very similar in meaning. Next, we merge *Ignore Greeting* (I8) with *Lack of Sociality* (I16, now E9), as the latter implies the first one. We merge *Repetition* (I15) and *Self-Contradiction* (I13) into a new error type, *Conversationality* (E7), since we observed both very rarely and only in situations that the system had lost the thread of the conversation. We also observed instances of incorrectly conveyed attributes (slots) that are not accounted for in the original taxonomy. To address this, we introduce the *Attribute Error* error type (E4).

4.2 User Response Type Taxonomy

We observed five different patterns in user responses subsequent to errors in system utterances during this study, which are also reflected in the collected error-indicating sentences. We propose them as a new taxonomy for the annotation of such data in dialogs (Table 5).

User Response Type	Description
Ignore and Continue (UR1)	The user ignores the error and continues the conversation, e.g., <i>Okay. Let's leave it like that.</i>
Repeat or Rephrase (UR2)	The user repeats or rephrases their concern, e.g., <i>Actually, I wanted ...</i>
Make Aware with Correction (UR3)	The user makes the system aware of the error and provides information to address what is missing or wrong in its utterance, e.g., <i>No. I wanted you to ...</i>
Make Aware without Correction (UR4)	The user makes the system aware of the error without providing additional information, e.g., <i>You're wrong.</i>
Ask for Clarification (UR5)	The user asks for clarification, e.g., <i>Are you sure? Is it really that ...</i>

Table 5: User Response Type Taxonomy.

Among these, we find that UR2, UR3, and UR5 are likely to contain free-text human feedback, such as corrections, new knowledge, or response alternatives.

5 Automatic Filtering for Potentially Relevant Dialogs

Since our study in Section 4 indicated that errors in system utterances are rare, we use Sentence-

Transformer (Reimers and Gurevych, 2019) to facilitate the process of filtering the remaining dialogs of each datasets for potentially relevant ones, i.e., dialogs with user responses similar to the collected error-indicating sentences.

For each dataset, we decompose every dialog into turns (alternating utterances), extract the user response, and segment it into sentences. Next, we pair these sentences with each of the error-indicating sentences and use a pretrained Sentence-Transformer based on MPNet (Song et al., 2020) to calculate their cosine similarity (see Appendix C for implementation details). We consider a dialog to be potentially relevant if at least one of these pairs has a cosine similarity $\geq 50\%$. Table 6 presents the sizes of the filtered subsets in comparison to the original datasets.

Dataset	Task-Oriented			Open-Domain		Know-Grounded
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Original Dialogs	8,438	16,000	6,235	10,907	60,000	22,311
Filtered Dialogs	4,936 (58.5%)	5,824 (36.4%)	421 (6.76%)	974 (8.9%)	15,960 (26.6%)	1,689 (7.57%)

Table 6: Size comparison between the filtered subsets and the original datasets. The numbers in brackets show the ratio of relevant dialogs to the original dataset sizes.

With 58.5%, MWOZ (Budzianowski et al., 2018) contains most of the potentially relevant dialogs. PC (Zhang et al., 2018) and WoW (Dinan et al., 2019) have the smallest number of such dialogs (8.9% and 7.57%, respectively). Overall, only 25% of the data is potentially relevant, i.e., contains at least one user response that is similar to one of those observed in Section 4. Hereinafter, we refer to these dialogs as **filtered dialogs**. We provide a sentence-level analysis in Appendix D.

6 Statistical Analysis

In this section, we conduct a statistical analysis of the distribution of error and user response types and their relations in the dialogs of the datasets examined. For this, we manually annotate 555 of the filtered dialogs (100 from each dataset, if available) with error and user response types, using the taxonomies proposed in Section 4.1 and 4.2. To avoid bias from our filtering procedure in Section 5, we also consider 600 randomly selected dialogs (100 from each dataset) that were not identified during this process (similarity $< 50\%$) in this analysis. Hereinafter, these dialogs are referred to as **random dialogs**. In Section 6.4, we assess the

impact of our filtering procedure on this analysis.

Overall, we manually annotate 1,155 dialogs with error and user response types. For annotation, we always consider the entire dialog (the context).

6.1 Error Type Distribution

Overall, we identified 188 errors across all dialogs. Table 7 shows the distribution.

Dataset	Task-Oriented			Open-Domain		Know-Grounded
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Filtered Dialogs	8/100	3/100	2/95	6/71	92/100	19/89
Random Dialogs	2/100	0/100	5/100	2/100	46/100	3/100

Table 7: The number of errors in comparison to the number of dialogs considered in this analysis for each dataset.

As expected, the filtered dialogs contain a larger number of errors (130 overall) compared to the random dialogs (58 overall), especially for open-domain and knowledge-grounded dialogs, such as SFC (Hancock et al., 2019) and WoW (Dinan et al., 2019).

Dataset	Task-Oriented			Open-Domain		Know-Grounded
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)
Ignore Question (E1)	1 (10.0%)	-	1 (14.3%)	1 (12.5%)	67 (48.5%)	-
Topic Trans. Error (E6)	-	-	-	1 (12.5%)	62 (44.9%)	4 (18.1%)
Factually Incorrect (E5)	-	2 (66.6%)	-	1 (12.5%)	3 (2.1%)	13 (59.1%)
Ignore Expect. (E3)	2 (20.0%)	1 (33.3%)	1 (14.3%)	-	2 (1.4%)	1 (4.5%)
Ignore Request (E2)	3 (30.0%)	-	1 (14.3%)	-	-	-
Lack of Sociality (E9)	-	-	-	2 (25.0%)	3 (2.1%)	-

Table 8: The most common error types and their frequencies found in both the filtered and random dialogs. The number in brackets shows the ratio to all errors found for the respective dataset.

Table 8 shows the most common error types and their frequency for both the filtered and random dialogs, which already accounts for 172 of all identified errors⁴. In the case of open-domain dialogs, the most frequent error types are *Ignore Question* (E1) and *Topic Transition Error* (E6). This is particularly the case in the SFC dataset (Hancock et al., 2019), where we find the system utterances to be often out of context. In the case of task-oriented dialogs, *Ignore Request* (E2) and *Ignore Expectation* (E3) are the most common error types. We

⁴See A1 in Table 15 and 16, Appendix G.1, for an aggregated distribution of all errors and user responses.

observe these errors when requests are only partially processed, e.g., when the user requests to book a hotel room and a train, but the system only books the hotel room. Moreover, we find that there is only little variety in language in task-oriented dialogs, regardless of the number of tasks reflected in the dataset (see Appendix E for examples). In the case of WoW (Dinan et al., 2019), the knowledge-grounded dataset, the *Factually Incorrect* (E5) error is the most frequently observed error type.

6.2 User Response Type Distribution

Table 9 shows the distribution of user response types to errors in system utterances for both the random dialogs (R) and the filtered dialogs (F).

Dataset	Task-Oriented				Open-Domain				Know-Grounded			
	MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)	
	F	R	F	R	F	R	F	R	F	R	F	R
Errors	8	2	3	0	2	5	6	2	92	46	19	3
UR1	1	2	2	0	1	3	0	1	4	35	0	1
UR2	2	0	1	0	1	0	0	0	0	0	0	0
UR3	2	0	0	0	0	2	0	0	3	1	9	0
UR4	1	0	0	0	0	0	2	1	34	2	0	1
UR5	2	0	0	0	0	0	4	0	51	8	10	1

Table 9: User response types found in the analyzed dialogs. For the random dialogs, we highlight the user response types that are likely to contain free-text human feedback (Section 4.2) in bold green and the other ones in bold red.

As described in Section 4.2, UR2 (*Repeat or Rephrase*), UR3 (*Make Aware with Correction*), and UR5 (*Ask for Clarification*) are likely to contain free-text human feedback. In the case of the filtered dialogs, we find that UR3 and UR5 are more often observed in open-domain and knowledge-grounded dialogs, such as SFC (Hancock et al., 2019) or WoW (Dinan et al., 2019). UR2 is only rarely observed, and only in task-oriented dialogs. However, UR1 (*Ignore and Continue*) is also frequently observed, especially in SFC. For randomly selected dialogs, this is the most frequent user response type (it occurs 42 times).

6.3 Relation Between Error and User Response Types

Figure 1 illustrates the relation between the most common error types (the 172 errors presented in Table 8) and user response types (Table 9) in both the filtered and random dialogs.

We find that UR1, UR4, and UR5 are the most frequently observed user response types, particu-

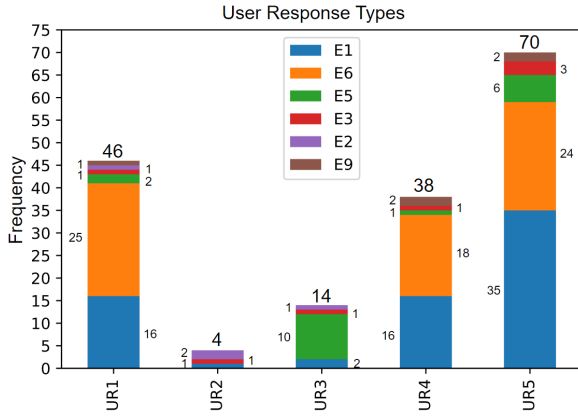


Figure 1: Illustration of the relations between frequent error (E-values) and user response types (UR-values) in both the filtered and random dialogs. The numbers above the bars are the total number of errors for each user response type. The numbers to the left and right of each bar indicate the portion of the respective error type (see color coding).

larly in the case of *Ignore Question* (E1) and *Topic Transition Error* (E6), which is mostly observed in open-domain datasets (Table 8). Along with UR3, UR5 is also a frequent response type in the case of *Factually Incorrect* (E5), which is mostly observed in WoW (Dinan et al., 2019). UR2 is only rarely observed. It sometimes occurs in the context of *Ignore Request* (E2) and *Ignore Expectation* (E3), which are mostly found in task-oriented dialogs.

6.4 Impact of Automatic Filtering

As Table 7 shows, a total of 188 dialogs was identified to contain errors in this analysis. 130 of them were found in filtered dialogs and 58 in random dialogs (the ones that were missed by our automatic filtering procedure in Section 5). Considering this at the level of user response types (Table 9), 46 of these 58 errors were ignored by users or did not provide any additional information (UR1 or UR4, the ones marked in bold red in the table), meaning that they are irrelevant because they do not contain free-text human feedback. For the remaining 12 missed errors (UR3 or UR5, the ones marked in bold green), we find that they are not reflected in the set of 67 error-indicating sentences used for filtering. Although this limits the effectiveness, we find that our procedure for automatic filtering itself has no negative impact on the results of our analysis, but rather improved annotation efficiency. An approximated recall of 0.72 supports this assumption. The recall was approximated with respect to the ratio between the size of the filtered subsets and

the original datasets⁵ (Table 6) and only considers the 12 missed relevant errors (when considering all missed errors, the recall is 0.35). We provide a more detailed analysis in Appendix F.

7 Evaluation and Experiments

In this section, we use the manually annotated dialogs from Section 6 in a human evaluation to assess the impact of our modifications to the Integrated Error Taxonomy (Higashinaka et al., 2021). We also use these dialogs to investigate the performance impact of errors in system utterances and subsequent user responses as additional input signals for response generation in three SOTA language generation models, including GPT-2 (Radford et al., 2019), LLAMA (Touvron et al., 2023), and Flan-T5 (Chung et al., 2022).

7.1 Integrated Error Taxonomy – Evaluation

To evaluate the impact of our modifications to reduce ambiguity and address missing error types in the Integrated Error Taxonomy (Higashinaka et al., 2021), we perform a human evaluation. We asked nine experts with NLP background and sound English skills to annotate 600 dialogs from those that were manually annotated by us in Section 6 (300 from both the filtered and random dialogs, 50 per dataset) with error and user response types using our modified Integrated Error Taxonomy (Section 4.1) and proposed user response type taxonomy (Section 4.2)⁶. Each of the dialogs was then assigned to two of these experts and thus annotated three times in total (including our own initial annotation). For comparison, we mapped all annotations back to the original Integrated Error Taxonomy. For merged error types, we asked the annotators for a second assessment using the original taxonomy. Table 10 shows the inter-annotator agreement (IAA) calculated using Krippendorff’s Alpha (Krippendorff, 2004)⁷ and summarized by human-human and human-bot dialogs.

In the case of human-human dialogs, the overall agreement is rather low. This also applies to the

⁵For calculating the recall, we randomly sampled 25% of the 555 annotated filtered dialogs, but considered all 600 random dialogs to reflect the proportions from Table 6. We repeated the sampling a thousand times and averaged the recall.

⁶We provide more details about the annotators (and a more detailed analysis of the results, including edge cases) in Appendix G. Please refer to Appendix H for the annotation guidelines.

⁷We use the Python library `annotation_analysis` for this, last accessed on 15. September 2023.

Annotation	Dataset	Ours		Theirs	
		HH	HB	HH	HB
Error Type	Filtered	0.16	0.91	0.02	0.89
	Random	0.17	0.40	0.16	0.39
User Response Type	Filtered	0.06	0.48	-	-
	Random	0.01	0.40	-	-

Table 10: Inter-annotator agreement using the Integrated Error Taxonomy (Higashinaka et al., 2021) (*Theirs*) and our modified version (*Ours*).

user response types (what was to be expected, as they directly depend on the error type annotations). During our analysis in Section 6, we found that in human-human dialogs, participants tend to suggest disagreement in a friendly manner, which complicates the identification of errors. This is different for human-bot dialogs, where participants tend to provide direct and clear feedback, such as *You’re wrong*. We attribute the low agreement for the annotations in human-human dialogs to this observation. Nonetheless, using our modified Integrated Error Taxonomy improves IAA over the original one in all cases. This is most obvious in the case of the human-human filtered dialogs, where it improves IAA by 0.14 points. A detailed analysis revealed that this is mainly due to (1) the condensed number of abstract error types, e.g., we merged ambiguous error types such as *Ignore Proposal* and *Ignore Request*, and (2) the newly added error types, such as *Factually Incorrect*, which were not covered in the original taxonomy.

7.2 Impact in Response Generation

In the following, we investigate the performance impact of including errors in system utterances and the subsequent user responses as additional input signals in response generation. For this experiment, we consider three different SOTA language generation models: GPT-2 (Radford et al., 2019), LLAMA (Touvron et al., 2023), and Flan-T5 (Chung et al., 2022). For GPT-2 and Flan-T5, we use the large variants. For LLAMA, we use the 7B variant⁸.

Experimental Setup We use the dialogs annotated in Section 6 for this experiment. In a first step, we use the 967 dialogs without error and user response type annotations to train baseline models in the task of response generation. Next, we train the baseline models using the 188 error and user response type annotated dialogs and include

⁸We use the pretrained models available in the [Hugging-face Model Hub](#) (last accessed 30. September 2023) for these experiments.

the annotated data as additional input signals. For error types, we include the respective system utterance (*Error Text*). For user responses, we include the respective user utterance (*User Response*). As evaluation metrics, we use word-overlapping F1-Score (following Xu et al. (2023) with FITS) and BLEU (Papineni et al., 2002). We provide more details, including the baseline results, in Appendix I.

Results Table 11 shows the results⁹. We find a large performance gap between Flan-T5 (Chung et al., 2022) and the other models. A detailed analysis revealed that both GPT-2 (Radford et al., 2019) and LLAMA (Touvron et al., 2023) generate reasonable and fluent responses, but mostly deviate from the target sequence. That aside, including user responses as an additional input signal improves the results over the other configurations, including *None*, for both Flan-T5 and GPT-2. For LLAMA, additionally using the error text improves the results over the other configurations.

Model	None		Error Text		User Response		Both	
	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU
Flan-T5	55.1	35.4	54.6	35.3	55.5	35.8	55.5	35.7
GPT-2	19.8	2.9	20.8	2.07	20.9	2.9	18.7	2.8
LLAMA	21.9	4.9	24.5	10.1	21.9	4.0	23.4	8.5

Table 11: Experiments with errors in system utterances and subsequent user reactions as additional input signals. For each model, the best performing configuration is highlighted. *Both* includes both feedback signals as additional input signal. *None* was just continually trained on the 188 dialogs, without including the feedback.

Overall, our results support the findings from recent works on learning from free-text human feedback in that including user responses to errors in system utterances is beneficial (Xu et al., 2023; Ung et al., 2022). However, we also show that including the error itself can have a positive impact.

8 Discussion

The goal of this work was to investigate the type and frequency of errors in system utterances and subsequent user responses included in the datasets examined to assess their extendibility with annotations for learning from free-text human feedback. We found that this mostly depends on whether the dialogs are human-human or human-bot. In human-human dialogs, we find that humans rather suggest

⁹In comparison to the baseline results (Appendix I), we see a performance drop in all configurations. We attribute this to the small number of annotated data and the varying quality.

disagreements in a very polite way instead of accusing the partner of a mistake (see Appendix J for examples). Accordingly, there is only little free-text human feedback available that could be used for learning (Section 6.2 and 6.3). Therefore, it might be hard and ineffective to extend these datasets with annotations for learning from such data. This is different in human-bot dialogs, where humans often react harshly and accusingly to errors in system utterances, resulting in more direct feedback. However, we also found that it depends on the dialog type. In general, we find that open-domain and knowledge-grounded dialogs contain a larger number of errors and user responses that are likely to contain free-text human feedback, making them more suitable for this purpose (Section 6.1).

Using the manually annotated dialogs from Section 6, our experiments in Section 7.2 suggest that including user responses to errors in system utterances has a positive impact in response generation, which supports the findings from recent works on including free-text human feedback (Xu et al., 2023; Ung et al., 2022). Additionally, our results suggest that including the error-annotated system utterance itself can have a positive impact. From our point of view, distinguishing between user response types could be an interesting alternative to binary signals, such as user satisfaction (Hancock et al., 2019) or thumbs-down (Shuster et al., 2022), as an indicator of an error in a system utterance. However, the dialogs annotated in Section 6 do not provide enough such data for a thorough analysis that also takes into account the different types of user responses. Therefore, we leave this as a research question for future work. Our human evaluation in Section 7.1 shows that our proposed taxonomies may serve as a promising starting point to obtain the necessary annotations, although they may not cover all possible error and user response types.

9 Conclusion

In this work, we examined the dialogs of six datasets from various types, including MultiWoZ, SGD, BABI, PersonaChat, Wizards-of-Wikipedia, and the human-bot split from the Self-Feeding Chatbot, for errors in system utterances and the types of subsequent user responses to assess their extendibility with annotations for learning from free-text human feedback. Our results show that this largely depends on whether the dialogs are

human-human or human-bot, and whether they are task-oriented, open-domain, or knowledge-grounded. We found that human-bot dialogs, contain more errors in system utterances that are addressed with free-text human feedback in subsequent user responses, especially in the case of open-domain and knowledge-grounded dialogs. Therefore, it might be feasible to extend these datasets with the needed annotations to support research into methods for learning from free-text human feedback, e.g., by taking advantage of the recent developments in synthetic data generation. We also used the insights gained during this process to propose a new user response type taxonomy and a modified Integrated Error Taxonomy for the annotation of free-text human feedback. Our experiments show that including errors from system utterances and subsequent user responses has a positive impact in response generation.

10 Limitations

The majority of our evaluation was done manually. Therefore, with respect to the original dataset sizes, we only consider a small fraction of the data in our study. It might be possible that our results would have been clearer when we would have considered more dialogs for the collection of error-indicating sentences. However, our analysis shows that errors found in the randomly selected dialogs are mostly ignored by the user, i.e., the user does not provide free-text human feedback that could be used for learning. Thus, as far as we are concerned, this does not limit the meaningfulness of our results.

Regarding dataset selection, our corpus study (and its results) have only limited expressiveness for knowledge-grounded dialog datasets, since we only consider one of such datasets in our study, Wizards-of-Wikipedia (Dinan et al., 2019). However, this does not affect the relevance of our work, as there are already free-text human feedback annotated datasets available, e.g., FITS (Xu et al., 2023), and we considered a representative number of datasets from other dialog types for which there is a lack of publicly available feedback-annotated datasets, such as task-oriented dialogs.

The taxonomies used in this work are also subject to limitations. In the case of the modified Integrated Error Taxonomy, our results show that it improves agreement across different dialog types. However, its abstract error types might limit application for specific use cases, e.g., for a more

fine-grained consideration of different types of social errors. Moreover, it reflects only error types observed in the datasets examined. The same applies to the user response type taxonomy.

11 Acknowledgments

This work has been funded by the LOEWE Distinguished Chair *Ubiquitous Knowledge Processing* (LOEWE initiative, Hesse, Germany) and the European Union under the Horizon Europe grant № 101070351 (SERMAS).

References

- Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. [Learning end-to-end goal-oriented dialog](#). In *International Conference on Learning Representations*.
- 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.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. [Deep Reinforcement Learning from Human Preferences](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- 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 Zhao, Yanping Huang, Andrew 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](#).
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *International Conference on Learning Representations*.
- Laila Dybkjaer, Niels Ole Bernsen, and Hans Dybkjaer. 1996. [Grice incorporated: Cooperativity in spoken dialogue](#). In *COLING 1996 Volume 1: The 16th International Conference 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.
- Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. [A repository of conversational datasets](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, 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.
- 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 *International Conference on Learning Representations*.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. [Soda: Million-scale dialogue distillation with social commonsense contextualization](#).
- Klaus Krippendorff. 2004. [Reliability in content analysis](#). *Human Communication Research*, 30(3):411–433.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Sebastian Möller, Klaus-Peter Engelbrecht, and Antti Oulasvirta. 2007. [Analysis of communication failures for spoken dialogue systems](#). In *INTERSPEECH 2007, International Speech Communication Association (ISCA)*.
- 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.

- Sunghyun Park, Han Li, Ameen Patel, Sidharth Mudgal, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. 2021. [A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational AI systems](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#).
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Abigail See and Christopher Manning. 2021. [Understanding and predicting user dissatisfaction in a neural generative chatbot](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. [Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage](#).
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tiejun Liu. 2020. [MPNet: Masked and Permuted Pre-training for Language Understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. [SaFeR-Dialogues: Taking feedback gracefully after conversational safety failures](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.
- Mathilde Veron, Sophie Rosset, Olivier Galibert, and Guillaume Bernard. 2021. [Evaluate on-the-job learning dialogue systems and a case study for natural language understanding](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023. [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.
- 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.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. [Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models](#).

A Integrated Error Taxonomy – Details

In this section, we describe the Integrated Error Taxonomy as proposed by Higashinaka et al. (2021). In principle, they differentiate between

form violation and *content violation*. The form violation usually represents errors that oppose some kind of meta criteria, e.g., the form of language or the ignorance of social norms. In contrast, content violations refer to, e.g., inconsistent or redundant utterances, or other things that might cause a dialog breakdown. Content violation is hereinafter abbreviated as *CV* (form violation as *FV*). Furthermore, they generally refer to *utterances*, while we refer to *system utterance*, as this is evident from their examples and simplifies understanding (from our perspective).

A.1 Utterance-Level

Utterance-level errors typically expose language-generation deficiencies of the system.

- **Uninterpretable (I1), FV** – The system’s utterance does not contain recognizable words or is just a text fragment, e.g., the utterance just consists of the words *of yours*.
- **Grammatical Error (I2), FV** – The system’s utterance is grammatical incorrect, e.g., *he will attend yesterday*.
- **Semantic Error (I3), CV** – The combination of predicates and arguments does not result in any meaning, e.g., *I feel painting*.
- **Wrong Information (I4), CV** – The given information is clearly wrong, e.g., *Meat Loaf was the songwriter of Jim Steinman* (it was vice versa).

A.2 Response-Level

Response-level errors describe errors in system utterances, e.g., in responses to user questions or requests.

- **Ignore Question (I5), FV** – The system utterance ignores the user’s question, e.g., *U: What will the weather be like tomorrow? S: Look at these beautiful flowers...*
- **Ignore Request (I6), FV** – The system ignores the user’s request to do something, e.g., *U: Please book me a room in the west part of town. S: I like sports.*
- **Ignore Proposal (I7), FV** – The system ignores the user’s proposal to do something, e.g., *U: Let’s talk about movies. S: I like pizza.*

- **Ignore Greeting (I8), FV** – The system ignores the user’s greeting, e.g., *U: Hello, my friend! S: What do you want from me?*
- **Ignore Expectation (I9), CV** – The system utterance does not have the expected content but the user’s intention has been successfully conveyed, e.g., *U: Do you have a favorite football team? S: Yes.* (the system should tell the user the name of his favorite football team).

A.3 Context-Level

Context-level errors refer not only to the local context, i.e., adjacent pairs of user utterance and system utterance, but to a broader (sometimes global) context.

- **Unclear Intention (I10), FV** – The system utterance is on the current topic, but the intention is unclear, e.g., *U: What is your favorite movie? S: I like Daddy’s Home 2. U: Oh, yes! Mel Gibson is so good in this role... S: The Evil Dead is so bloody...*
- **Topic Transition Error (I11), FV** – The system suddenly switches to another topic without explanation, e.g., *U: I really enjoyed Feynman’s physics lecture. So enlightening! S: Well, what is your favorite food?*
- **Lack of Information (I12), FV** – The system utterance misses important information, such as subject, object, or modifier, e.g., *U: It’s a beautiful day, isn’t it? S: I hate.* (object is missing).
- **Self-Contradiction (I13), CV** – The system utterance contradicts what it said before, e.g., *S: I’m from Boston, Massachusetts. U: Oh, wow! I’m from Bartlesville, Oklahoma! S: Really? I’m from Seattle, Washington!*
- **Contradiction (I14), CV** – The system utterance contradicts what has been said by the user, e.g., *U: I like strawberries. S: Oh, why don’t you like strawberries?*
- **Repetition (I15), CV** – The system repeats something it has already said before (this includes repeated questioning), e.g., *S: I look forward to the concert. U: Oh, which concert? S: I look forward to the concert.*

A.4 Society-Level

Society-level errors refer to errors that indicate that the system lacks etiquette or common sense.

- **Lack of Sociality (I16), FV** – The system utterance lacks consideration towards the user and contains hate speech or is in general disrespectful, e.g., *U: Yesterday I went to the federal employment agency. S: Okay. Are you incapable of finding a job on your own?*
- **Lack of Common Sense (I17), CV** – The system utterance contains content that differs from the opinion of the majority, e.g., *S: I like to eat chicken raw. U: Oh, dear... You have to cook it first!*

B Error-Indicating Sentences And Phrases

In this section, we present the collected error-indicating sentences along with phrases.

1. Not really like fandoms, haha Just anything online that people make. (**Phrase:** *not really like*)
2. It is not saturday. (**Phrase:** *it is not*)
3. That doesn't make sense. (**Phrase:** *doesn't make sense*)
4. That makes no sense. (**Phrase:** *makes no sense*)
5. You should put some more things together." (**Phrase:** *you should*)
6. You shouldn't be! (**Phrase:** *you shouldn't*)
7. What do you mean by that?" (**Phrase:** *what do you mean*)
8. What are you talking about? (**Phrase:** *what are you talking about*)
9. It's so important for young people to have diverse interest and develop a wide range of skills, don't you think? (**Phrase:** *don't you think*)
10. I don't know what you're talking about. (**Phrase:** *don't know*)
11. What does that have to do with computer games? (**Phrase:** *what does that have to do with*)
12. Sorry I meant to say for the cat litter. (**Phrase:** *sorry i meant to say*)
13. That didn't have anything to do with school. (**Phrase:** *didn't have anything to do with*)
14. You do not make sense with your response. (**Phrase:** *your response*)
15. That's not what I asked you. (**Phrase:** *not what i asked*)
16. I dont understand. (**Phrase:** *don't understand*)
17. How do you mean? (**Phrase:** *how do you mean*)
18. I don't care about price. (**Phrase:** *i don't care*)
19. You're not answering the questions. (**Phrase:** *you're not answering*)
20. Like I said before I'm not one to read an actual newspaper but I do like reading opinion and political articles. (**Phrase:** *like i said before*)
21. You're not very helpful Help Desk. (**Phrase:** *not very helpful*)
22. Are you sure that there are no hotels on the west side of town? (**Phrase:** *are you sure*)
23. I didn't say anything was scary. (**Phrase:** *i didn't say*)
24. I wouldn't know this. (**Phrase:** *i wouldn't know this*)
25. That sounds too low. (**Phrase:** *too low*)
26. I'm great, but thats off topic. (**Phrase:** *that's off topic*)
27. No, I think when people shape their beards in different ways is really interesting as well! (**Phrase:** *no, I think*)
28. Your doing it wrong my friend. (**Phrase:** *you're doing it wrong*)
29. What are you saying? (**Phrase:** *what are you saying*)
30. At least you have that then. (**Phrase:** *at least you have*)
31. That doesn't answer my question. (**Phrase:** *that doesn't answer*)

32. I am too old to hike I am in my seventies. (**Phrase:** *i am too old*)
33. You aren't staying on topic at all. (**Phrase:** *not staying on topic*)
34. Off the subject, I am thinking of cutting my hair. (**Phrase:** *off the subject*)
35. I'm not ready to book just yet. (**Phrase:** *i'm not ready*)
36. That's not what I asked you. (**Phrase:** *i asked you*)
37. Dude not cool. (**Phrase:** *dude not cool*)
38. I'd really like a 4 star. (**Phrase:** *i'd really like*)
39. Thats nonsense." (**Phrase:** *thats nonsense*)
40. Actually, I apologize no need to book, I was just gathering information." (**Phrase:** *i apologize*)
41. I never said I needed one. (**Phrase:** *i never said i*)
42. No I dont think so. (**Phrase:** *no i dont think*)
43. I didn't mention anything about clowns. (**Phrase:** *i didn't mention*)
44. That is odd for alaska. (**Phrase:** *that is odd*)
45. Not sure what that means? (**Phrase:** *not sure what that means*)
46. It can be what? (**Phrase:** *it can be what*)
47. You should learn! (**Phrase:** *you should learn*)
48. Umm, what? (**Phrase:** *umm, what*)
49. You think so? (**Phrase:** *you think so*)
50. No a park is a place and not a person, (**Phrase:** *and not*)
51. Why do you say that? (**Phrase:** *why do you say that*)
52. I guess I should have asked that first. (**Phrase:** *i should have asked*)
53. I said lets talk about sports. (**Phrase:** *i said lets talk about*)
54. You're being annoying is whats happening. (**Phrase:** *you're being annoying*)
55. You could have stated the goods. (**Phrase:** *you could have stated*)
56. Who was talking about color? (**Phrase:** *who was talking about*)
57. That doesn't really matter. (**Phrase:** *doesn't really matter*)
58. It's actually a 1939 movie that was adapted from a novel written earlier. (**Phrase:** *it's actually*)
59. I don't believe a piano is a stringed instrument. (**Phrase:** *i don't believe*)
60. That's not relevant. (**Phrase:** *that's not relevant*)
61. Check again. (**Phrase:** *check again*)
62. You're wrong. (**Phrase:** *you're wrong*)
63. That doesn't have to do with track. (**Phrase:** *that doesn't have to do with*)
64. Instead could it be in Madrid? (**Phrase:** *instead could it*)
65. I would prefer in Bombay. (**Phrase:** *i would prefer*)
66. No, I don't like that. (**Phrase:** *i don't like that*)
67. No, this does not work for me. (**Phrase:** *this does not work*)

C Automatic Filtering – Implementation

To implement the automatic filtering (Section 5) we use PyTorch (Paszke et al., 2019), the Transformers library (Wolf et al., 2020), and the pretrained *all-mpnet-base-v2* Sentence-Transformer¹⁰. It is based on MPNet (Song et al., 2020) and finetuned on a large corpus of sentence pairs from multiple tasks and domains, e.g., Yahoo Answers (Zhang et al., 2015) and Reddit Comments (Henderson et al., 2019), using a contrastive objective. It is a 12-layer Transformer model with a vocabulary size of 30,527 words that calculates the cosine similarity between two sentences in a 768-dimensional dense vector space.

Our compute infrastructure consists of one Tesla V100-SXM3 GPU (with 32 GB memory) and it

¹⁰Model page in the HuggingFace Model Hub, last accessed 11. September 2023.

took an average of 76 mins to run automatic filtering on one dataset.

D Automatic Filtering – Sentence-Level Analysis

As described in Section 5, we filter on sentence-level for similar user responses. Figure 2 illustrates the ranges of similarity between the sentences extracted from the user utterances and the error-indicating sentences, i.e., 50% – 60%, 60% – 70%, 70% – 80%, 80% – 90%, 90% – 100%. It reflects the share in identified phrases from each of the datasets (see Table 3). Most of the phrases were identified in SFC (Hancock et al., 2019). Only a small amount of phrases came from the other datasets which might be the reason for the clusters in the lower ranges.

E Task-Oriented Dialogs – Examples

In Section 6.1, we say that the dialogs of task-oriented datasets considered in our work are of little variety in language and flow. This is most obvious in case of BABI (Bordes et al., 2017) which we want to show using the following example:

1. **USER:** good morning
1. **SYSTEM:** hello what can i help you with today
2. **USER:** may i have a table with british cuisine in a cheap price range
2. **SYSTEM:** i'm on it. where should it be
3. **USER:** madrid please
3. **SYSTEM:** how many people would be in your party
4. **USER:** for eight please
4. **SYSTEM:** ok let me look into some options for you
5. **USER:** actually i would prefer in paris
5. **SYSTEM:** sure is there anything else to update
6. **USER:** instead could it be for four people
6. **SYSTEM:** sure is there anything else to update
7. **USER:** instead could it be in a moderate price range
7. **SYSTEM:** sure is there anything else to update

8. **USER:** no

8. **SYSTEM:** ok let me look into some options for you

This dialog is prototypical for BABI. The first few turns (the first four in this case) contain the original user request. Then, the user asks for changes (turn five to eight). Overall, the dialogs are very scripted. Users are mostly clear and distinctive in their responses. They contain only little variety in language and flow, which makes them seem very unrealistic. We do not observe a change in this regard across tasks.

The following dialog is from SGD (Rastogi et al., 2020):

1. **USER:** Our family is planning for a vacation trip. Please can you find a good rental car for a pleasant trip?
1. **SYSTEM:** At what time do you need the car?
2. **USER:** I can pick up the car at 15:00
2. **SYSTEM:** From when onwards you are looking for the rental car?
3. **USER:** I need the car from 11th of March.
3. **SYSTEM:** So, how long you need the car?
4. **USER:** I might need it till the 13th of this month.
4. **SYSTEM:** Okay! So, where would be your pick up location?
5. **USER:** I can pick up the car in Paris on 9th of this month.
5. **SYSTEM:** Okay! There is a nice car, Chevrolet Cruze, a Standard car is available for pick up on March 9th at Charles de Gaulle International Airport. Do you wish to rent this car?
6. **USER:** Do you have any other cars for I need the car till 10th of this month.
6. **SYSTEM:** Yup! We've 3 more cars. Do you wish to rent Honda Accord, which is a Standard car and its ready for pickup on March 9th from Charles de Gaulle International Airport?
7. **USER:** Okay. I go for the Honda Accord.

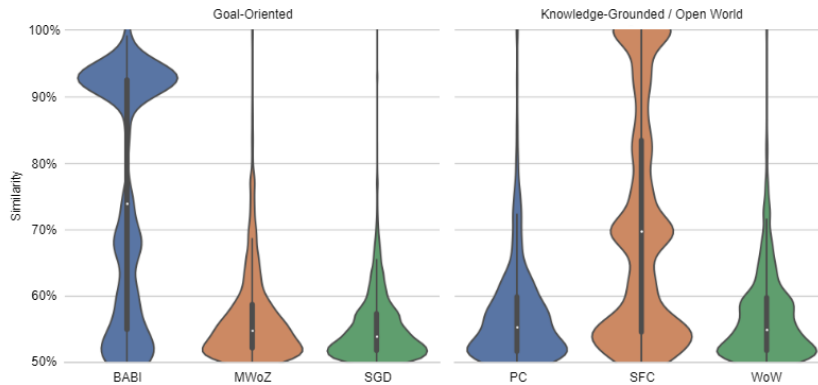


Figure 2: Illustration of the distribution of dialogs with potential real-world error scenarios across datasets.

The dialog looks much more natural, which might be due to its type. While BABI consists of human-bot dialogs, SGD consists of human-human dialogs. However, the flow is very similar. Up until turn five, the user’s request is refined. Then, the system comes up with a suggestion which the user denies (for a couple of turns) by asking for something different. At some point, user’s agree or disagree on the system’s suggestions. Then the dialog ends.

The following dialog is from MWoZ (Budzianowski et al., 2018):

1. **USER:** Hi, I could use help with my planning. I am trying to find a place to stay.
1. **SYSTEM:** Ok, to help us find the right place for you lets talk about what you would like. Do you have a price range you prefer? Or an area of town?
2. **USER:** I would like it to be in the south and have a rating of 3 stars.
2. **SYSTEM:** We have a moderately priced guest-house and an expensive hotel that match your requirements.
3. **USER:** Does either offer free parking?
3. **SYSTEM:** They both offer free parking and internet.
4. **USER:** Okay can you please book it for 7 people for 4 nights starting on Tuesday?
4. **SYSTEM:** Which one? The guesthouse or the hotel?
5. **USER:** The hotel, please.

5. **SYSTEM:** Reference number is : BS0ODSGJ. Is there anything else?

6. **USER:** Yes, I also need a train from stevenage to cambridge. I want to leave after 10:15 on tuesday.

6. **SYSTEM:** I have train TR0517 that leaves at 11:54. Would that be something that would work for you?

7. **USER:** That’s perfect. Could I get the train ID and travel time?

7. **SYSTEM:** The travel time is 49 minutes.

8. **USER:** Great, that’s all I need, thanks so much for your help! Have a great day!

8. **SYSTEM:** You’re very welcome! Thank you! Bye!

Like in case of SGD, the dialog looks very natural. However, in contrast to BABI and SGD, the dialogs consist less of refinements, but more of multiple tasks as shown in this example. First, the user asks for a hotel in a city, then he asks for transportation to that city. Sometimes, he also asks for locations with entertainment. While the first request is usually completed (the booking of a room in this case), the second request is usually about gathering information (the user just asks for the train number, departure and travel time, but not for booking a seat).

F Effectiveness Of Automatic Filtering – A Detailed Analysis

For the statistical analysis in Section 6, we consider 20 dialogs from each similarity range, i.e.,

50% – 60%, 60% – 70%, 70% – 80%, 80% – 90%, 90% – 100% (if available, see also Appendix D) for each dataset examined. As the data in the upper ranges (80% – 100%) is scarce in case of WoW (Dinan et al., 2019), PC (Zhang et al., 2018), and BABI (Bordes et al., 2017), the filtered dialogs consists only of 555 dialogs (instead of 600 like the randomly selected dialogs). Table 12 shows the errors annotated for the statistical analysis with respect to the similarity ranges identified by automatic filtering (meaning that each dialog contains at least one user response with a sentence identified to be similar to at least one error-indicating sentence in this similarity range). *Overall* (O) represents the number of dialogs randomly sampled from the respective similarity range, and *Error* (E) represents the number of dialogs identified in our manual analysis to contain an error in a system utterance.

Dataset	Task-Oriented						Open-Domain				Know-Grounded		
	MWoZ (HH)		SGD (HH)		BABI (HB)		PC (HH)		SFC (HB)		WoW (HH)		
Overall / Error	O	E	O	E	O	E	O	E	O	E	O	E	
Filtered Dialogs	90% - 100%	20	2	20	2	17	0	6	2	20	20	9	4
	80% - 90%	20	2	20	1	18	0	5	2	20	20	15	9
	70% - 80%	20	1	20	0	20	0	20	0	20	19	20	4
	60% - 70%	20	1	20	0	20	2	20	1	20	18	20	2
	50% - 60%	20	2	20	0	20	0	20	1	20	15	20	0
Overall	100	8	100	3	95	2	71	6	100	92	89	19	
Random Dialogs	100	2	100	0	100	5	100	2	100	46	100	3	

Table 12: Identified errors in all datasets across similarity ranges.

Overall, only 58 dialogs of the randomly selected ones (9.6%) contain errors. In the case of automatic filtering, we observe 130 of such cases. Therefore, automatic filtering shows to facilitate the process of identifying errors in system utterances. Even if the number of identified errors is overall low, most errors are identified in the range of 60% – 100%, excluding the densest section in case of MWoZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), PC and WoW, 50% – 60% (see also Figure 2).

G Inter-Annotator Agreement – Detailed Analysis

This section gives more insights on the inter-annotator agreement. All additional annotators that participated in this study were non-native speakers. They were experts from our lab with sound English skills and NLP background. We did not select them based on specific criteria; they participated voluntarily. Accordingly, they were not paid extra for this, since they did the annotations

during their working hours. For annotation, we did not use any specific tool. We provided the annotators with dialogs in json format and asked them to do the annotations directly in the respective files. See Section H for the annotation guidelines. Table 13 shows the inter-annotator agreement for each dataset using our modified Integrated Error Taxonomy.

Dataset	Task-Oriented			Open-Domain		Know-Grounded	
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)	
Error Type	Filtered	0.01	0.0	1.0	0.51	0.81	0.12
	Random	0.55	0.01	-0.01	0.09	0.80	0.02
User Res. Type	Filtered	0.04	0.0	0.23	0.16	0.72	0.04
	Random	0.05	0.0	0.0	0.01	0.79	-0.02

Table 13: Inter-annotator agreement for each dataset.

In the case of human-human dialogs, the overall agreement is rather low (except for PersonaChat (Zhang et al., 2018)). We find that errors are hard to identify in these dialogs, as humans rather suggest disagreements instead of accusing the partner of a mistake. This is also reflected in the user response type agreement since it depends on the error type annotation. However, PersonaChat seems to be different (according to Table 8). We attribute this to the dialog type, which is open-domain, where we find that humans react harshly and accusing to errors in system utterances, resulting in more direct feedback that is easier to identify.

Dataset	Task-Oriented			Open-Domain		Know-Grounded	
	MWoZ (HH)	SGD (HH)	BABI (HB)	PC (HH)	SFC (HB)	WoW (HH)	
Error Type	Filtered	-0.10	0.0	1.0	0.26	0.80	-0.09
		(-0.11)	(-0.0)	(-0.0)	(-0.25)	(-0.01)	(-0.21)
Random	0.55	0.01	-0.01	0.09	0.80	0.0	
	(-0.0)	(-0.0)	(-0.0)	(-0.0)	(-0.01)	(-0.02)	

Table 14: Inter-annotator-agreement for the Higashinaka et al. (2021) taxonomy.

Table 14 shows the inter-annotator agreement for each dataset using the original Integrated Error Taxonomy as proposed by Higashinaka et al. (2021). Using this taxonomy deteriorates the inter-annotator agreement. This is most obvious in case of MWoZ (Budzianowski et al., 2018) and PC (Zhang et al., 2018), which are both human-human datasets. A detailed analysis revealed that this is mostly due to over-specialized error types which were merged in our modified taxonomy, such as *ignore expectation* and *ignore request*, I9 and I6 in the original taxonomy (Table 2). Another reason are the newly added error types, such as *factually*

incorrect, E5 (Table 4), which were not covered in the original taxonomy, but occur in the dialogs.

G.1 Edge Cases

Table 15 shows the aggregated error type distribution (error type annotation from both the filtered and random subsets). *A1* denotes the authors’ annotations done for the statistical analysis in Section 6.

	Task-Oriented			Open-Domain			Knowledge-Grounded											
	MwZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
Ignore	1	2	2	-	-	1	1	2	1	1	2	5	67	64	66	-	1	3
Question (E1)	3	3	3	-	-	-	1	-	-	-	-	-	1	9	6	-	-	-
Ignore	2	3	3	1	-	1	1	1	1	-	-	-	2	1	-	2	-	-
Expect. (E3)	3	1	-	-	-	-	4	-	3	-	2	3	-	3	-	1	5	1
Attribute Error (E4)	-	2	-	2	-	-	-	4	-	1	2	-	3	1	-	13	1	1
Factually Incorrect (E5)	-	-	1	-	-	-	-	-	-	2	2	10	62	58	58	4	-	1
Topic Trans. Error (E6)	1	-	2	-	-	1	-	-	1	1	1	1	-	-	2	1	-	3
Convers. (E7)	-	12	-	-	-	-	-	-	-	-	-	1	-	2	2	-	13	-
Unclear Intention (E8)	-	-	-	-	-	-	-	-	-	2	1	4	3	2	1	-	-	-
Lack of Sociality (E9)	-	-	1	-	-	-	-	-	-	1	2	2	-	-	-	1	-	1
Lack of Com. Sense (E10)																		

Table 15: Error types in both the filtered and random dialogs.

Overall, the distribution is very broadly spread. However, in most cases, it seems like at least two annotators agree. There are only a few outliers where there is a large deviation, i.e., unclear intention (E8 in Table 4) in case of MwZ (Budzianowski et al., 2018) and WoW (Dinan et al., 2019), topic transition error (E6) in case of PC (Zhang et al., 2018), factually incorrect (E5) and attribute error (E4) in case of WoW. For example, attribute error is defined as an error type that rather addresses task-oriented dialogs, but annotator two found it five times in the WoW dataset. During our analysis, we found that factually incorrect would have described these cases more accurately. In the case of unclear intention in WoW and MwZ, we found that annotator two marked some cases as errors that are actually not necessarily errors. The same applies to the factually incorrect errors in BABI (Bordes et al., 2017) (which consists of task-oriented dialogs). In the case of PC, we found that topic transition error is in most cases the most obvious error type, and in our opinion, annotator three was right in most of the cases.

In summary, we find that deviations are primarily the result of (1) how the annotators interpret the descriptions of the error types (based on their experience), and (2) biases in the data. The former could probably be addressed by more examples in the annotation guidelines. The latter is a bit more

difficult. In these cases, a multi-step annotation process could be useful, where annotators mark errors they are not sure about to be discussed before they are finally annotated.

	Task-Oriented			Open-Domain			Knowledge-Grounded											
	MwZ (HH)			SGD (HH)			BABI (HB)			PC (HH)			SFC (HB)			WoW (HH)		
Annotator	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
UR1	3	8	5	2	-	3	4	2	3	1	2	23	39	40	36	1	6	2
UR2	2	9	3	1	-	-	1	3	1	-	-	1	-	1	-	-	-	-
UR3	2	3	3	-	-	-	2	1	-	-	-	-	4	3	-	9	7	5
UR4	1	3	1	-	-	-	-	1	1	3	5	2	36	37	35	1	-	-
UR5	2	-	-	-	-	-	-	-	-	4	5	-	59	59	64	11	7	3

Table 16: User response types in both the filtered and random dialogs.

Table 16 shows the aggregated distribution of user response types. What stands out here are the differences in the situations where users ignore the error situations (UR1 in Table 5) in the case of PC (annotator three) and WoW (annotator two). However, we found no pattern in the underlying error situations. We suspect that this is due to the nature of human-human dialogs (PC and WoW consist of human-human dialogs). As we have pointed out before, humans behave more cautiously and politely in human-human dialogs, while in contrast they clearly point out errors in human-bot dialogs (see Appendix J for examples).

H Annotation Guidelines

H.1 Annotation Task

Learning from textual user feedback, i.e., free-form text feedback that expresses user dissatisfaction, describes new knowledge (new concepts), provides corrections or alternative responses, is increasingly important to continuously improve dialog systems. As of today, only few datasets provide annotation for such data. However, there are various dialog datasets publicly available – from a variety of domains and for all kinds of tasks. As an annotator, you are provided with two different taxonomies, an error type taxonomy, and a user response type taxonomy, and two different types of dialogs, conspicuous dialogs and cold dialogs. Your task is to annotate the dialogs using the given taxonomies. The conspicuous dialogs are identified to contain conspicuous sentences in user utterance that address an error in the previous system response. You have to decide (1) whether or not this conspicuous sentence correctly addresses an error in the previous system response, (2) which error type is represented in the previous system response, and (3) how the user reacted. The cold dialogs do not

contain such information. In this case, your task is to identify error situations and user response types in the dialogs just given the taxonomies.

H.2 Dialog Format

To avoid bias, we provide the dialogs in a unified format without the original task-specific annotations. Here's an example of a conspicuous dialog:

```
{
  "unique_id": "PMUL0121.
  json_multiwoz_train",
  "id": "PMUL0121.json",
  "turns": [
    {
      "utterance": "Hi there! Can you
        give me some info on
        Cityroomz?",
      "response": "Cityroomz is
        located at Sleeperz Hotel,
        Station Road in the Centre
        area. It is a moderatley
        priced hotel with a 0 star
        rating. Would you like to
        book a room?"
    },
    {
      "utterance": "Yes please. I
        need it for 7 people for 3
        nights starting on tuesday.",
      "response": "How many days
        would you like to book it
        for?"
    },
    {
      "utterance": "I would like to
        book it for 3 nights
        beginning on Tuesday.",
      "response": "Ok, Your rooms
        have been booked. Your
        reference number is
        CYJDAR7R . Is there
        anything else I can help
        you with today?"
    },
    {
      "utterance": "Yeah, I need a
        restaurant in the west and
        with expensive pricing.",
      "response": "There are 9
        expensive restaurants in
        the west area. Do you
        prefer a certain type of
        food?"
    },
    {
      "utterance": "No, I don't have
        a preference, but I need
        the phone number and
        postcode.",
      "response": "I recommend
        Graffiti, the number is
        01223277977 and the
        postcode is cb30lx. Would
        you like me to book a table
        for you?"
    }
  ]
}
```

```
},
{
  "utterance": "No, right now, I
    just need the infor. Thanks
    !",
  "response": "Can I help you
    with anything else today?"
},
{
  "utterance": "That is all.
    thanks",
  "response": "Thank you for
    using our system."
}
],
"conspicuous": [
  {
    "phrase": [
      "actually, i apologize no
        need to book, i was just
        gathering information.",
      " i would like to book
        it for 3 nights
        beginning on tuesday."
    ],
    "confidence": 0.593,
    "turn": 2
  }
],
"annotations": [
  {
    "turn": 2,
    "annotation": {
      "error_type": "E2",
      "comment": "the system
        misses intent/slots. the
        user already said that
        he need it for three
        nights",
      "error": "C1",
      "user_response": "B3"
    }
  }
]
}
```

Each dialog consists of a unique id, an id, and its turns. Conspicuous is an array. The first value is an error-indicating phrase, a phrase that was identified to express user dissatisfaction in the utterance of the corresponding turn. The second value is the value from an utterance of this dialog that was identified to be similar to this error-indicating sentence. Confidence represents the similarity. Dialogs with multiple conspicuous values are possible. The annotations list has an entry for each conspicuous phrase. Please add your annotations here. In comment, you can share your thoughts with us.

Here's an example for an cold dialog:

```
[
  {
    "dialog": "p2 cats are like
      cartoons. p1 that's cool ,
      whats your favorite food ? p2
      pizza. p1 ni hao . as my
```

```

    father says . you must have
    great plans ahead ? p2 yes, i
    plan to be a success.",
    "error": "C2",
    "error_type": "",
    "user_response": "",
    "comment": "",
    "turn": "",
    "phrase": "",
  },
  ...
]

```

The structure is a bit different. All cold dialogs are provided in one large json file, and the dialogs themselves maintain the structure of the original dataset. In this case, it is an dialog from the human-bot split of the Self-Feeding Chatbot (p2 represents the system, p1 represents the user). There are two additional fields here: turn and phrase. If you by chance find a phrase that indicates dissatisfaction in the user's response to a system's error, please add phrase and turn to these fields.

H.3 Taxonomies

H.3.1 Error Type Taxonomy

This is the taxonomy for the field error type.

```

[
  {
    "id": "E1",
    "name": "Ignore Question",
    "description": "The response ignores a user's question.",
    "example": "User: 'How is the weather in Berlin?', System: 'Have you ever heard of Bobby Jean?'"
  },
  {
    "id": "E2",
    "name": "Ignore Request",
    "description": "The response ignores a user's request to do something.",
    "example": "User: 'Can you book me a room for two please?', System: 'Look at these flowers ... Aren't they beautiful?'"
  },
  {
    "id": "E3",
    "name": "Ignore Expectation",
    "description": "The response does not fulfill the user's expectation.",
    "example": "User: 'Can you book me a room at The Inn for two nights starting tomorrow? Can you also provide me with the address?', System: 'Absolutely !'"
  },
]

```

```

{
  "id": "E4",
  "name": "Attribute Error",
  "description": "The response suggests that the system did not get the attributes/slots right.",
  "example": "User: 'Can you book me an expensive restaurant in the west of town?', System: 'Sure. I would recommend The Breakfast Club. It is a very cheap and nice location.'"
},
{
  "id": "E5",
  "name": "Factually Incorrect",
  "description": "The response contains information that is factually incorrect.",
  "example": "User: 'Who is the current president of the United States of America?', System: 'Abraham Lincoln.'"
},
{
  "id": "E6",
  "name": "Topic Transition Error",
  "description": "The response transitions to another / a previous topic without reasonable explanation.",
  "example": "User: 'I think I have a cold.', System: 'I really love pears. So juicy!'"
},
{
  "id": "E7",
  "name": "Conversationality",
  "description": "The response indicates that the system lost track, e.g., it repeats previous responses (without asking for missing information) or contradicts itself.",
  "example": "System: 'Man is shaped by his environment.', User: 'Yes. But does that justify this tone?', System: 'Man is shaped by his environment.'"
},
{
  "id": "E8",
  "name": "Unclear Intention",
  "description": "The response suggests that the user's intent was not successfully conveyed.",
  "example": "User: 'Can you book me a flight to London?', System: 'Yes. What do you want for dinner?'"
},
]

```

```

{
  "id": "E9",
  "name": "Lack of Sociality",
  "description": "The response lacks consideration of social standards, e.g., greetings, is toxic or disrespectful.",
  "examples": "User: 'Hi! My name is Jochen Schweitzer!', System: 'Sure. How can I help you?'"
},
{
  "id": "E10",
  "name": "Lack of Common Sense",
  "description": "The information in the responses opposes the opinion of the majority.",
  "example": "User: 'You have to work hard for success.', System: 'Hmm... No. What is to be, will be.'"
}
]

```

H.3.2 User Response Taxonomy

This is the taxonomy for the field user response.

```

[
  {
    "id": "UR1",
    "short": "The user ignores the error and continues the conversation.",
    "description": "The user simply continues and does not draw the system's attention to the error.",
    "example": "-"
  },
  {
    "id": "UR2",
    "short": "The user repeats or rephrases his/her concern.",
    "description": "The user repeats or rephrases his originally concern.",
    "example": "'Can you book a restaurant for two for tonight?' vs. 'Can you book a table for two for tonight?'"
  },
  {
    "id": "UR3",
    "short": "The user makes the system aware of the error and provides a correction.",
    "description": "The user makes the system aware of the error and provides information to address what is missing

```

```

or wrong in its utterance. ",
"example": "'No, I didn't want you to book a table. I just wanted the address!'"
},
{
  "id": "UR4",
  "short": "The user makes the system aware without providing a correction.",
  "description": "The user makes the system aware without providing additional information",
  "example": "'No. You're wrong.'"
},
{
  "id": "UR5",
  "short": "The user asks for clarification.",
  "description": "The user is puzzled and asks for clarification, e.g. the system suddenly switches to another topic or mixed concepts up.",
  "example": "'What do you mean?'"
}
]

```

I Hyperparameters and Baseline Experiments

Hyperparameters All baseline models were trained for five epochs. For the experiment using erroneous dialogs, we trained the models for ten epochs. 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 ¹¹ (Loshchilov and Hutter, 2019) in PyTorch. Except for LLAMA (Touvron et al., 2023), we fully-finetuned all models. For LLAMA, 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.

Results Table 17 shows the results of our baseline experiments using word-overlapping F1-Score and BLEU (Papineni et al., 2002).

Compared to the results of our experiment using erroneous dialogs (Table 11), both F1-Score and BLEU are higher in case of GPT-2 and LLAMA. We attribute this to the comparatively bad quality of the erroneous dialogs. The majority of these

¹¹AdamW in the Pytorch documentation (last accessed 30. September 2023).

	F1	BLEU
Flan-T5	56.3	35.7
GPT-2	47.7	36.6
LLAMA	57.9	42.1

Table 17: Results of our baseline experiments.

dialogs comes from the SFC (Hancock et al., 2019) dataset, where most errors are due to out-of-context system utterances (see Section 6.1) and were ignored by users (see Section 6.2).

J Human-Human Dialogs – Examples

We find that it hard to identify errors in human-human dialogs, as the tone is in principle mostly polite, even if one of the partners is wrong. The following dialog is an example from WoW (Dinan et al., 2019):

1. **USER:** I adore photography even though I'm bad at it. It was never a good idea to give me access to a DSLR camera...
1. **SYSTEM:** Well, photography is a science. A little practice and you'd get it down.
2. **USER:** You think so? I think most people, myself included, would consider in an art. But I suppose it's true that there is a lot of technical detail to it.
2. **SYSTEM:** I personally do it as a hobby but many do in businesses, film ect. In my opinion it's always art. i love it because it can capture what your heart see's. If you angle it just right

Here, one partner claims that photography is science. However, it's commonly considered as a visual art. Instead of directly correcting him, the other person rather suggests that the system might be wrong. The following example is from PC (Zhang et al., 2018):

1. **USER:** Hi, how it is going today?
1. **SYSTEM:** It is going great today; everyday is great in my ford truck.
2. **USER:** Oh. I am eating a piece of fried chicken.
2. **SYSTEM:** Right on. I do not allow fried chicken in my ford truck.
3. **USER:** What do you do for a living?

3. **SYSTEM:** I work on ford trucks. Do you have a ford truck?
4. **USER:** No , i drive a toyota.
4. **SYSTEM:** I am so sorry. You should get yourself a ford truck. Mine is black .
5. **USER:** I am not. I love my toyota .
5. **SYSTEM:** Well okay. How about a ford car? They are just as good.
6. **USER:** Nope, hardcore toyota family. My wife drives a 4runner.
6. **SYSTEM:** Why do they call it a 4runner?
7. **USER:** Because it will run forever.
7. **SYSTEM:** Oh, I thought it was because 4 wheeler was taken.
8. **USER:** Oh. That does not make much sense.
8. **SYSTEM:** Are you kidding me? Do you know any jokes?

In the second to last turn, one partner makes a joke that the other does not understand. The response is quite polite. He informs the partner that his statement does not make much sense to him.