Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?

Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
{mahed.mousavi,gabriel.roccabruna,giuseppe.riccardi}@unitn.it

Abstract

Human Evaluation (HE) of automatically generated responses is necessary for the advancement of human-machine dialogue research. Current automatic evaluation measures are poor surrogates, at best. There are no agreed-upon HE protocols and it is difficult to develop them. As a result, researchers either perform non-replicable, non-transparent and inconsistent procedures or, worse, limit themselves to automated metrics. We propose to standardize the human evaluation of response generation models by publicly sharing a detailed protocol. The proposal includes the task design, annotators recruitment, task execution, and annotation reporting. Such protocol and process can be used as-is, as-a-whole, in-part, or modified and extended by the research community. We validate the protocol by evaluating two conversationally fine-tuned state-of-the-art models (GPT-2 and T5) for the complex task of personalized response generation. We invite the community to use this protocol - or its future community amended versions - as a transparent, replicable, and comparable approach to HE of generated responses[1].

1 Introduction

Early attempts to evaluate automatic Natural Language Generation (NLG) models using human judges dates back to before the appearance of end-to-end models (Jones and Galliers, 1995; Coch, 1996; Lester and Porter, 1997). However, due to the expensive requirements such as training skilled annotators and the time-consuming nature of this evaluation, automatic metrics became the common evaluation criteria in several NLG tasks. Metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) have been used to evaluate the model performance in machine translation and automatic summarization tasks as inexpensive and rapid evaluations. After observing the reliability of these metrics for the task they are designed for (if applied correctly), they have been used to evaluate the models in other tasks such as response generation. However, several studies have shown that currently available automatic metrics can not be good candidates for evaluating a generated response (Liu et al., 2016; Sai et al., 2022); these criteria co-relate poorly with human judgement and are inadequate since the generation is subject to trivial factors such as coherency, fluency and grammatical structure, as well as non-trivial factors such as appropriateness, engagement, and user acceptance.

Human Evaluation (HE) is still the necessary approach to evaluate the generated responses (Smith et al., 2022). With the development of crowdsourcing annotation platforms, conducting an HE task is less expensive and more feasible than early methodologies. Nonetheless, little attention has been given to the assessment of the design of HE task. Due to the lack of an agreed upon and standard protocol, HE tasks have been performed while suffering from nontransparent procedures, non-replicable and incomparable results, and unclear resource allocations.

In this work, we propose to standardize the experimental methodology for human evaluation for response generation models. We present a detailed protocol to the community for this task, in order to increase the comparability, replicability, and interpretability of such evaluations among works and domains. All the required steps and materials to conduct a HE in a transparent and extendable way (including task design, annotator recruitment, task execution, and annotation reporting) are described and shared with the community. The proposed protocol is domain-agnostic, language-independent, and open to be extended to different versions and standards. We invite the community to not only utilize this protocol but also to improve and extend it into referable and version-able standards for the

[1] Link to the protocol and materials Repository

HE task. In order to validate the proposed protocol, we evaluate two conversationally fine-tuned state-of-the-art models for the Italian language, based on GPT-2 (De Mattei et al., 2020) and T5 (Sarti and Nissim, 2022), for the task of response generation in personal dialogues using knowledge grounding.

2 Literature Review

Earlier attempts to evaluate dialogue systems by human judges considered user satisfaction as the evaluation criterion (Walker et al., 1997). Despite the introduction of automatic metrics and a research direction aiming to better the metrics used for the evaluation of dialogue models (Zhang et al., 2019; Huang et al., 2020; Mehri and Eskenazi, 2020), Human Evaluation (HE) is still the gold standard for assessing the qualities of a generated response and a generative model.

While the importance of the proper evaluation of a dialogue model using human judges is well-established in the community, how to perform such evaluation is still an unsolved question (Smith et al., 2022). As an outcome, countless HE tasks have been presented and conducted in this domain, resulting in non-comparable and non-replicable results. Dialogue systems have been evaluated with different granularity (turn-level vs dialogue-level), different evaluation policies (single-model vs pairwise-model, candidate-ranking vs. winner-selection) and in different modalities (interactive vs static) (Smith et al., 2022). The ambiguities in HE tasks conducted so far have also been studied by Belz et al. (2020), where the authors focused on disentangling the characteristics of already conducted HE tasks to increase the interpretability and comparability of the evaluations and results. Further inconsistency in the evaluations includes the ambiguity in the criterion name, i.e. two criteria with the same name assess two different qualities in different works, whereas the same quality has been named with various terms among works (Howcroft et al., 2020). In addition to the aforementioned works, this naming inconsistency can also be found in the grounded generation literature (Zhang et al., 2020; Wang et al., 2020; Huang et al., 2021; Hedayatnia et al., 2020; Zhang et al., 2020) where a criterion with the same name refers to two different qualities and presents different definitions among works.

An important factor for reproducing any crowd-sourcing experiment is reporting the details related to that experiment and its settings. This issue has been studied by Ramírez et al. (2021), where the authors identify the properties that researchers have to provide to facilitate the reproducibility of any crowd-sourcing experiments. The same problem has been studied specifically for HE experiments by Howcroft et al. (2020) where the authors identify the lack of reporting crucial details, and other issues such as high levels of variation among the evaluation procedures. Howcroft et al. (2020) further stress the need for a standard and coherent experimental design and terminology for the task of HE in the community.

3 Proposed Human Evaluation Protocol

We propose to standardize the Human Evaluation (HE) experiments through a referable and replicable protocol to address the problems of non-comparability and inconsistency in the literature. Considering the complexity of designing and executing such evaluations, we unfold the task into four main steps in order to study and analyze the crucial aspects at each step. We aim to maximize the reliability and replicability of the evaluation while minimizing the task difficulty and complexity. Our proposed protocol consists of four executive steps, i.e. 1) Task Design; 2) Annotator Recruiting; 3) Task Execution; and 4) Reporting.

3.1 Step 1: Task Design

The first step is to design the evaluation task, which can be characterized by the two aspects of evaluation and annotation characteristics. Defining these characteristics clearly and transparently is paramount to achieve replicability and comparability among works and models.

3.1.1 Evaluation Characteristics

As the initial step, definition of the evaluation characteristics of the task include the evaluation granularity, quality dimensions to evaluate and their definitions, the questions to be asked to the annotators, and the annotations format.

Granularity The evaluations conducted in the literature can be categorized into two levels of granularity as dialogue-level, where the model is evaluated at the end of a complete dialogue, and turn-level, where the model is evaluated based on its output for a specific turn in the dialogue. Recent works indicate that the turn-level evaluation is more fine-grained since it captures errors such as contradictions and response repetitions (Smith et al., 2022). Turn-level evaluation can be further categorized as absolute (single-model, or rating) or comparative (winner-selecting, or ranking). In this protocol, we evaluate the models at the turn-level; and in order to avoid biasing the annotators with the quality of other candidates which may result in an unintentional pick-the-best response, we evaluate the candidates using the absolute setting (i.e. presenting one candidate per time for each dialogue history). In this way, the performance quality of each model is evaluated independently and we can obtain a model-specific list of limitations and error signals. Furthermore, the ground truth turn is also provided as a response candidate to the annotators, representing a point of reference.

Quality Dimensions We include four criteria in this version of the protocol, based on the most common errors and qualities for an end-to-end response generation model. Nevertheless, the proposed protocol can be extended to other criteria and quality dimensions. The proposed criteria and their definitions are as follows;

- **Appropriate** whether the proposed response candidate makes sense with respect to the dialogue history; and to investigate if it is a proper continuation of the given dialogue (thus coherent).
- **Contextual** whether the proposed response candidate contains references to the dialogue context (thus not generic); and to investigate whether the response refers to non-existing or contradicting information (such as model hallucination).
- **Listening** whether the speaker of the proposed response is following the dialogue with attention (note that generic responses are also indicating that the speaker is not following the dialogue).
- **Correct** whether the response candidate is correct considering the grammar, syntax and structure of the response.

Questions One of the important details, which is usually missing in the evaluation reports in the literature, is the formulation of the questions the annotators are prompted for the quality of the responses. The questions must be designed in a clear and neutral form in order to avoid any possible bias while addressing the important factors evaluated by each criterion. We present the questions designed to evaluate the responses in each dimension in the Appendix, Section A (The protocol can be expanded to other dimensions used by adding the corresponding criteria and questions).

Decisions For each criterion, the annotators are asked to select an answer from a 3-point Likert scale modeled as positive (eg. Correct, Appropriate), negative (eg. Not Correct, Not Appropriate), and *"I don't know"*. The purpose of the third choice, *"I don't know"*, is to avoid forcing non-deterministic and error-prone judgements on one of the other two options. That is, the non-expert annotator (in some cases nor the expert annotator) may not be able to make a deterministic decision due to the residual and inevitable ambiguity of the annotation task.

Explanations In order the obtain better insights into the capabilities and limitations of the models, we ask the annotators to explain their judgement by pointing out possible errors or rightness of a response. The explanation is asked for three of the criteria (listening is excluded) and mostly when the response is negatively evaluated or the annotator is not sure ("*I don't know.*"). In order to introduce the minimum amount of cognitive workload to the task, the annotators are asked to explain their judgement for each response right after evaluating a response candidate, through predefined options to select from, and/or free text. The list of predefined explanation options to select from and the cases for which the explanation is asked is presented in Table 1.

3.1.2 Annotation Characteristics

Another principal aspect of HE experiments is the annotation characteristics. Despite the importance of this aspect and its influence on the resulting quality, little attention is given to the careful design of the HE annotation task.

We can model the annotation task as the interactions of the human (in our setting the annotator) with a task system (the evaluation). From the beginning of the task, the annotator tends to create a Mental Model of the task according to the properties and information she/he is presented to (Moray, 1998). One of the main causes of issues in such settings is the gap between the user's and the designers' mental model (Norman, 1988; Xie et al., 2017). Furthermore, studies show high levels of cognitive workload in a task reduce the humans' ability to retrieve and exploit knowledge while reducing the mental workload helps to reduce the frequency of errors (Leveson, 2016; Zenati et al., 2020; Ramírez et al., 2021). Therefore, it is necessary to carefully design the annotation task to

Quality Dimension	Annotators' Decision		Quality Sub-dimension
	Value	Explanation Options	
Appropriateness	Appropriate	□ ***"The proposed response is coherent with the dialogue context."***	*Coherence*
		□ Add free form text explanation	-
	Not Appropriate	□ ***"The proposed response is not coherent with the dialogue context."***	*Incoherence*
		□ Add free form text explanation	-
	I don't know	□ **Please Add free form text explanation (required)**	-
Contextualization	Not Contextualized	□ ***"The response is generic or does not contain any explicit or implicit reference to what it has been said in the dialogue context."***	*Genericness*
		□ ***"The response is not consistent with the information contained in the dialogue context."***	*Hallucination*
		□ Add free form text explanation	-
	I don't know	□ **Please Add free form text explanation (required)**	-
Correctness	Not Correct	□ ***"The response contains grammatical errors."***	*Grammaticality*
		□ ***"The response contains one or more parts that are repetitive."***	*Repetition*
		□ Add free form text explanation	-
	I don't know	□ **Please Add free form text explanation (required)**	-

Table 1: The explanation options provided to the annotators to support their decisions. The annotators can select predefined option(s) and/or write a free form text. Each explanation option refers to a sub-dimension which is used as interpretations for the result analysis. The sub-dimensions are not presented to the annotators.

ensure a controlled level of cognitive workload throughout the task and minimize the possibility of misunderstanding or ambiguity for the annotators by using well-explained guidelines, a simplified User Interface, and a clear annotation process.

Guidelines & Examples An important resource in crowd-sourcing annotation tasks is the guidelines, which have the objective to introduce the task to the annotator and instruct them about the process. The task guidelines and the examples must be written with a clear and simple structure in order to minimize possible ambiguities for the annotators and help them form a mental model in line with the one of the task designers. The examples should be carefully selected to point out the possible ambiguities and difficulties during the annotation and to help the workers get familiar with the task. Our task guidelines include an introduction to the task, the definition and description of each criterion and corresponding answer sets, as well as examples of various scenarios and annotations. The complete format of this version of our guidelines can be found on repository.

User Interface We designed and implemented a User Interface (UI) for the task of Human Evaluation, with the objective of an easy-to-use and intuitive platform that is extendable to other versions of the evaluation. A complete description of the UI is presented in Appendix, Section C.

Internal Pilots Internal pilots can provide reliable feedback about the difficulty/subjectivity of the task, the amount of time required to perform the task, and a threshold for the expected output quality of the task if done correctly. Internal pilots also help to detect and resolve possible ambiguity and issues in the task and its materials prior to the main task.

3.2 Step 2: Annotator Recruitment

After designing the task, we need to recruit the required number of annotators to perform the task. In most cases the annotation is done through crowd-sourcing. In that case, there are several aspects involved in the process of recruiting the crowd-workers that can affect the outcome quality including the sampling policy, the qualification, and the compensation.

Sampling In order to obtain reliable results, it is important to recruit the annotators from the correct target group. Selecting the annotators in the literature has been mostly conditioned by prerequisites such as location, language fluency, and level of education. Further, Mousavi et al. (2021) studied the impact of domain expertise in a domain-specific annotation task.

Qualification Karpinska et al. (2021) observed that when the annotators are sampled from workers in crowd-sourcing platforms, sampling conditions are not adequate as they may be fulfilled inappropriately (for instance the use of VPNs to fake a certain location). Therefore, in addition to the mentioned prerequisites, it is essential to set up a qualification task for the workers. The qualification task helps the task designers to filter out contributors with low-quality performance and helps the crowd-workers to get familiar with the main task and the UI.

Compensation Proper compensation is an important extrinsic factor that can affect the performance of crowd workers, and the time it takes for the job to be selected and worked on by the workers (Mason and Suri, 2012; Whiting et al., 2019; Ramírez et al., 2021). Therefore, it is crucial to estimate properly and fairly the time and complexity needed to complete the task and set a fair wage in order to ensure a proper compensation.

3.3 Step 3: Task Execution

The execution of the main task is subject to continuous control of the progress and quality. In this phase, the agreement level among the annotators can indicate whether the outcome quality is maintained throughout the task. Sudden drops or jumps in the agreement level can be due to unbalanced difficulty among batches, or a low-quality contributor. While the former should be addressed using stratified sampling when designing the task, Riccardi et al. (2013) observed that providing real-time feedback to the annotators helps them to recover their mistakes and improve their performance for the upcoming tasks.

3.4 Step 4: Annotation Reporting

Howcroft et al. (2020) highlights the lack of a standard for reporting the description and the results of HE experiments and points out the need for proper reporting of the evaluation details and results analysis. Furthermore, Ramírez et al. (2021) stresses the importance of reporting the crowd-sourcing experiment in a proper and standard way in order to facilitate replicability of the experiment and reproducibility of the results. In this protocol, we provide a checklist of aspects and elements that are necessary to be reported along with the final results in order to ensure a clear and transparent presentation of the protocol and possible outcomes. The characteristics of the task that should be reported are:

- Evaluation granularity (dialogue-level vs. response-level, comparative vs. absolute)
- Quality dimensions, their definitions, and corresponding questions
- Annotation format (item selection, free form text, ranking, rating, etc.)

While the details regarding the recruitment of the crowd-workers include:

- Sampling criteria, the description of qualification task and acceptance\rejection criterion
- Number of workers recruited

Besides the mentioned details, there are certain statistics related to the execution of the evaluation task and its final outcome that should be reported to increase the credibility of the results. These statistics include:

- Annotators participated in the study
- Samples annotated in the study
- Votes per each sample
- Inter-Annotator Agreement level & the metric used
- Workload allocated per annotator
- Demographic of the annotators
- Resource Utilization (time to perform the task, payment to the annotator, crowd-sourcing platform)

4 Validation of the Protocol

We validate the proposed protocol by evaluating two response generation models for the task of personal and grounded response generation. For this purpose, we fine-tuned two of the state-of-the-art Pre-trained Language Models (PLMs) for the Italian language. The first model fine-tuned is iT5 (Sarti and Nissim, 2022), which has the same architecture as T5 PLM (Raffel et al., 2020), pre-trained on a large Italian corpus. We used iT5-Base which consists of 12 layers per stack (encoder or decoder) with 220M parameters. The second model fine-tuned in this work is GePpeTto (De Mattei et al., 2020) which is a decoder-only autoregressive PLM based on GPT-2 small (Radford et al., 2019), for the Italian language. The model consists of 12 layers of decoder and byte-pair encoding, with 117M parameters.

The fine-tuning of the two models was done using the dataset of Follow-Up dialogues collected by Mousavi et al. (2021). This dataset is a collection of dyadic conversations about personal events and emotions the narrator has experienced while the listener tends to respond with personalized and helpful suggestions. The dialogues in this data set are based on a personal narrative about the same event and participant that the narrator has shared prior to the dialogue. We fine-tuned the models with and without grounding the generation on the corresponding narrative for each dialogue via the same approach used by Zhao et al. (2020). In our setting the knowledge selection module is not required since the correct narrative for each dialogue is deterministic.

iT5-Base was fine-tuned using AdaFactor optimizer (Vaswani et al., 2017) and early stopping wait counter equal to 3, with batch size and dialogue history window equal to 4. GePpeTto was fine-tuned using AdamW optimizer (Loshchilov and Hutter, 2017) and early-stopping wait counter equal to 3, with batch size and dialogue history window equal to 2. For fine-tuning the models 80% of the dataset was used, while 10% was used as the validation set for early stopping and parameter engineering and the rest of the data, unseen 10%, was used as the test set (the splits were sampled at dialogue level to ensure no history overlap among splits). The automatic evaluation of fine-tuned models is presented in Table 3.

4.1 Implementation of the HE Protocol

We implemented the proposed protocol to evaluate the performance of the two models by human crowd-workers.

Task Design We followed the Task Design step explained in subsection 3.1 closely. We then sampled 42 different dialogue histories from the fine-tuning test set (approximately 50%) for the evaluation (the length of histories varies from 2 to 4 turns) and sampled the responses of all models for each dialogue. We conducted two internal pilots using 5 dialogues with 3 internal experts (the experts were not involved in the design of the task), as well as 3 internal non-expert annotators. After each pilot, the feedback of both groups was collected and few refinements were made to the UI and the guidelines.

Using the feedback obtained from the internal pilots regarding the difficulty of the task and the amount it takes to annotate the samples, we prepared the annotation batches so that each batch consists of approximately 10 dialogue histories of 4 turns in average, with 3 response candidates (including the ground truth) to evaluate for the next turn. During the internal pilots, each batch of 5 dialogues took an average of 15 minutes for the non-expert annotators. Therefore, we set the average required time to 35 minutes and the maximum time possible to annotate a batch to 90 minutes, in order to factor in the possible lower pace of non-expert annotators.

Recruiting Crowd-worker We used Prolific crowd-sourcing platform[2], and selected the crowd-workers using the following prerequisites:

- **Location**: Italy
- **Gender Distribution**: Available to All
- **First Language**: Italian
- **Minimum Approval rate**: 95%
- **Minimum complete submissions**: 20 jobs
- **Education**: Available to all
- **Expertise**: Available to all

In addition to the sampling policy, the annotators were asked to perform a qualification task. The task consisted of evaluating the response candidates for 5 dialogues (same dialogues used in the internal pilots) in an identical setting to the main task. We considered the Inter-Annotator Agreement (IAA) of the internal non-expert annotators calculated by Fleiss' κ (Fleiss, 1971) as the threshold (0.21). In order to qualify each worker, we computed the agreement level among the internal annotators and the worker and if it was above the threshold, the worker was qualified for the main task.

Based on the workload and the estimated time required for the task, we set the wage as 4.67 pounds for 35 minutes, equal to 8 pounds per hour[3]. In this protocol, qualified crowd-workers were also paid for the qualification task.

4.2 Annotation Statistics

In total, 40 workers participated in the annotation task and 35 of them were qualified. The 42 samples to annotate were distributed in two batches of 11 and two batches of 10 samples. Each batch is annotated by 7 annotators and the annotators spent an average of 19 minutes for the qualification batch and 45 minutes for annotating the main batches. In addition to the decided compensations, one annotator was rewarded a bonus of two pounds since he/she informed us about an unexpected bug in the UI via email.

[2] Prolific: `https://www.prolific.co/`

[3] Prolific's Payment Principles mandates a fair and ethical payment to the workers with the minimum of 6 pounds (8 dollars) per hour. While deploying the study on the platform, the task owner is prompted with recommended payment level for the study, for which our payment of 8 pounds per hour was labelled as "Good".

Models	**Inter Annotator Agreement Level measured by Fleiss' κ**				
	Appropriateness	***Contextualization***	***Correctness***	***Listening***	**IAA per Model**
GePpeTto	0.27	0.14	0.64	0.15	0.32±0.10
+Knowledge	0.42	0.22	0.36	0.27	0.36±0.11
iT5-Base	0.24	0.19	0.06	0.18	0.27±0.04
+Knowledge	0.18	0.03	0.30	0.21	0.19±0.06
IAA per Dimension	0.30 ±0.10 **Fair**	0.15±0.05 **Poor**	0.41±0.20 **Moderate**	0.23±0.07 **Fair**	-

Table 2: The Inter-Annotator Agreement (IAA) level calculated by Fleiss' κ. The last row and last column represent the average IAA (and the standard deviation) per each of the criteria and each model, respectively. The low IAA on *Contextualization* indicates the high level of complexity and subjectivity in this criterion. In contrast, the moderate level of IAA is achieved over *Correctness* criterion, suggesting a lower level of subjectivity in the judgements.

Models	**Automatic Evaluation**		**Human Evaluation**			
	nll	***ppl***	**Appropriateness**	**Contextualization**	**Correctness**	**Listening**
Ground Truth	-	-	100.0%	97.62%	97.62%	97.62%
GePpeTto	2.76	15.84	66.67%	69.05%	83.33%	64.29%
+Knowledge	2.79	16.33	59.52%	57.14%	83.33%	57.14%
iT5-Base	2.05	7.79	66.67%	73.81%	100.0%	66.67%
+Knowledge	2.04	7.70	80.95%	80.95%	85.71%	76.19%

Table 3: The automatic and human evaluation outcome of the fine-tuned models. The results are obtained by majority voting. The evaluations indicate that grounding mostly improves the performance of iT5 Base, while it worsens GePpeTto's performance. Note that the perplexity can not be compared among models since the pre-training data and thus the vocabulary distributions are not identical.

During the execution of the task, we calculated the agreement between each pair of annotators using Cohen's kappa (Cohen, 1960) as well as the agreement among all annotators in the same batch using Fleiss' κ (Fleiss, 1971) metrics. We further calculated the agreement among all annotators on strong judgements, by removing items that were labelled as "I don't know." by at least one annotator. Despite little fluctuations in the agreement level, no low-quality contributions were detected and the agreement level on different batches was consistent throughout the evaluation.

Table 2 presents the average Inter Annotator Agreement (IAA) measured by Fleiss' κ. The agreement is calculated per each model and criterion in each batch (for the 7 annotators who annotated the batch) and averaged over all batches. The results indicate that *Contextualization* and *Listening* are the two criteria with the highest levels of subjectivity and complexity. In contrast, high IAA over *Correctness* suggests that it has been easier for the annotators to assess the grammatical and structural aspects of the response samples.

4.3 Evaluation Results

Table 3 presents the results of the HE based on the majority voting for each model. While the grounding generally improved the performance of iT5-Base, it worsened the performance of GePpeTto in all aspects. Nevertheless, it introduced grammatical and structural errors in iT5-Base output. Moreover, grounding did not improve GePpeTto to generate more contextualized responses. While grounded iT5-Base outputs were evaluated the highest among the models, there is still a huge gap to reach the quality of the ground truth. This matter shows the complexity of generating an appropriate and contextual response in personal dialogues.

Figure 1 represents the sub-dimension errors that the annotators selected to explain their negative votes on the response candidates. The explanation option corresponding to each error is presented in Table 1. The figure is obtained by considering all the votes of the annotators on every response sampled from the models (each response is evaluated by 7 annotators, thus 294 votes in total). Therefore, for instance, while iT5-Base achieves 100% of "*Correctness*" by majority voting, there are 7

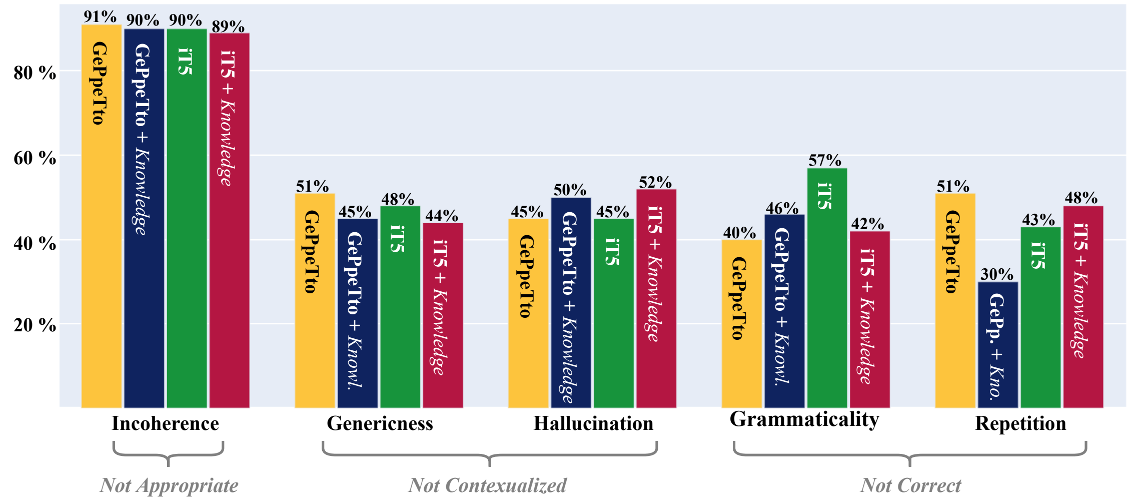


Figure 1: The sub-dimension errors selected by the annotators for the explanation of negative judgements in each criterion. Each bar represents the percentage of the times the error category (x-axis) was selected as the reason to reject the output of the corresponding model. The figure is obtained by considering all the votes (i.e. not majority voting). Note that the labels are not mutually exclusive.

cases (out of 294) where the annotators labelled it as "Not Correct"; the selected reason in 4 cases was grammatical error and in 3 cases a repetition in the response.

These results indicate that, regardless of the model, while grounding reduces the cases that a response is labelled as "*Not Contextualized*" due to being a *Generic* response, it increases the cases of *Hallucination* problem with almost the same proportion. Nevertheless, the percentage of cases where a response is labelled as "*Not Appropriate*" due to being *Incoherent* is not affected by the grounding technique and all models suffer from this error equally. Furthermore, we observe that grounding slightly increases the cases in which a response by GePpeTto is labelled as "*Not Correct*" due to errors related to *Grammaticality*, while it considerably reduces the cases of *Repetition* in such responses.

In addition to the pre-defined explanations, in a few cases the annotators also provided us with free-form explanations. Specifically, in 10% of the cases in which the model outputs were labelled as "*Not Correct*", the annotators provided us further explanations to indicate the exact grammatical error such as punctuation or subjunctive errors (Congiuntivo in Italian). In 5% of the times in which the model responses were considered "*Not Contextualized*" the annotators pointed out the exact part of the response which is mentioning a wrong event/participant, or is in contradiction to the dialogue history. Lastly, in 10% of the cases where the response candidate was evaluated as "*Not Appropriate*" the annotators provided explanations to highlight the exact segment of the response that is not right or is ambiguous.

5 Conclusion

While Human Evaluation is the necessary methodological step in the assessment of response generation models, there is a lack of a standard. This deficiency has resulted in often ambiguous, incomparable and non-replicable published experiments. In this work, we aim at addressing this problem by sharing a complete methodology for evaluating generated responses using human judges. We publish the first version of the protocol and all its materials to the community. The expectation is to engage them to utilize, extend, and complement this protocol into further versions and a transparent resource that can be publicly accessed. The ultimate goal is to engage the community to consider HE as an important topic of research. The complete protocol and supporting materials can be found in a public repository (including the guidelines, task design, the UI, and the analysis scripts).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation* *measures for machine translation and/or summarization*, pages 65–72.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *INLG*.

José Coch. 1996. Evaluating and comparing three text-production techniques. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421.

David M. Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *INLG*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240.

Xinxian Huang, Huang He, Siqi Bao, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Plato-kag: Unsupervised knowledge-grounded conversation via joint modeling. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 143–154.

Karen Sparck Jones and Julia R Galliers. 1995. Evaluating natural language processing systems: An analysis and review.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285.

James Lester and Bruce Porter. 1997. Developing and empirically evaluating robust explanation generators: The knight experiments. *Computational Linguistics*, 23(1):65–101.

Nancy G Leveson. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Neville Moray. 1998. Identifying mental models of complex human–machine systems. *International Journal of Industrial Ergonomics*, 22(4-5):293–297.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34.

Giuseppe Riccardi, Arindam Ghosh, SA Chowdhury, and Ali Orkan Bayer. 2013. Motivational feedback in crowdsourcing: a case study in speech transcription. In *INTERSPEECH*, pages 1111–1115.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*.

Yanmeng Wang, Wenge Rong, Jianfei Zhang, Yuanxin Ouyang, and Zhang Xiong. 2020. Knowledge grounded pre-trained model for dialogue response generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206.

Bingjun Xie, Jia Zhou, and Huilin Wang. 2017. How influential are mental models on interaction performance? exploring the gap between users' and designers' mental models through a new quantitative method. *Advances in Human-Computer Interaction*, 2017.

Marco A Zenati, Lauren Kennedy-Metz, and Roger D Dias. 2020. Cognitive engineering to improve patient safety and outcomes in cardiothoracic surgery. In *Seminars in thoracic and cardiovascular surgery*, volume 32, pages 1–7. Elsevier.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Appendix

A Questions

Dimension	Question	Answer Option	Option Definition
Appropriateness	*Is the proposed response candidate appropriate?*	Appropriate	The response makes sense and it can be the natural continuation of the shown dialogue context.
		Not Appropriate	The response does not make sense in the current dialogue context.
		I don't know	The candidate contains some elements which make sense with respect to the dialogue context, but some that do not.
Contextualization	*Does the proposed response contain references to the context of the dialogue?*	Contextualized	The candidate contains implicit or explicit references to the dialogue context.
		Not Contextualized	The candidate doesn't contain any reference to the dialogue context, or contains references that are incoherent with the dialogue context.
		I don't know	The response contains some references to the dialogue context, but contains other references that are not clear or relevant.
Listening	*In the proposed response candidate, how much do you think person A is listening to person B?*	Listening	Speaker A is listening with attention to speaker B and follows the dialogue.
		Not Listening	Speaker A seems not to pay attention to what speaker B is saying.
		I don't know	It is unclear if speaker A is listening to speaker B or not.
Correctness	*Is the proposed response grammatically correct?*	Correct	The response does not contain any type of grammatical or structural error, any repetitions, misspellings or any other types of error.
		Not Correct	The response contains some grammatical or structural errors such as, repetitions, misspelling, any other types of error.
		I don't know	It is hard to identify if the response contains errors or not.

Table 4: The questions and possible answer options presented to the annotators for the evaluation of the response candidates in this version of the protocol. The complete version of the option definitions is presented to the annotators in the guidelines.

B Model Response Evaluation

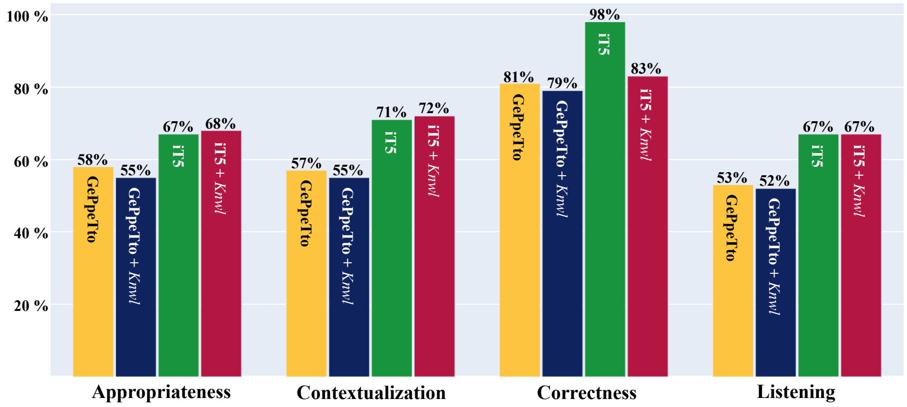
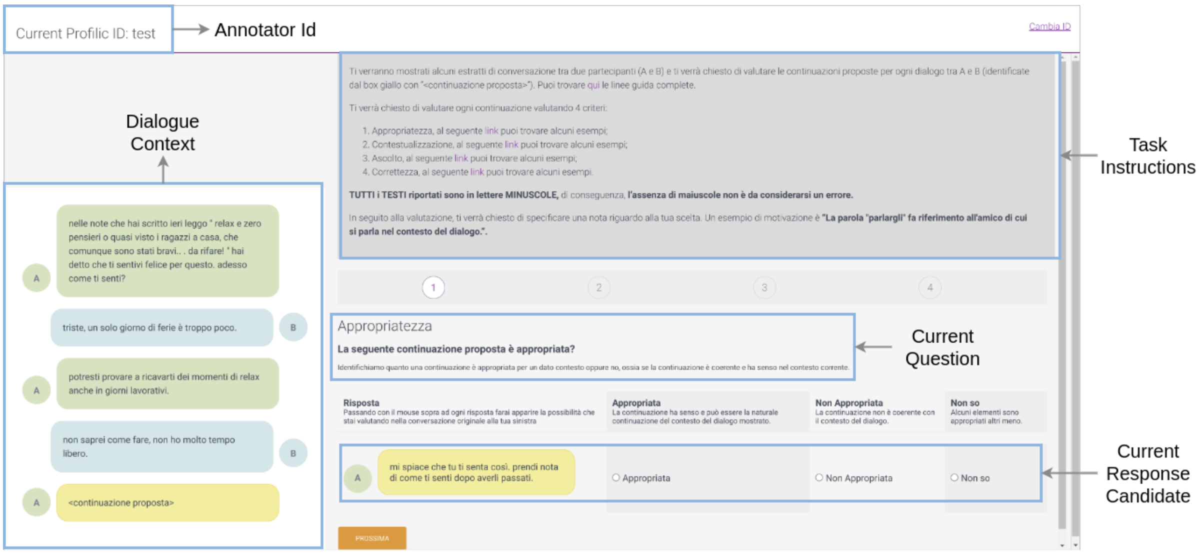


Figure 2: The human evaluation of the models in each criterion by considering all the votes (i.e. not majority voting). Each bar represents the percentage of the times the corresponding model was labelled positively by the criterion on x-axis. While Table 3 is obtained by majority voting, this figure is obtained by considering all the annotators votes on the response samples (i.e. not majority voting). iT5-Base variations outperform GePpeTto variations, regardless of the presence of grounding.

C User Interface



Current Profilic ID: test

Cambia ID

Ti verranno mostrati alcuni estratti di conversazione tra due partecipanti (A e B) e ti verrà chiesto di valutare le continuazioni proposte per ogni dialogo tra A e B (identificate dal box giallo con "<continuazione proposta>"). Puoi trovare qui le linee guida complete.

Ti verrà chiesto di valutare ogni continuazione valutando 4 criteri:

1. Appropriatezza, al seguente link puoi trovare alcuni esempi;
2. Contestualizzazione, al seguente link puoi trovare alcuni esempi;
3. Ascolto, al seguente link puoi trovare alcuni esempi;
4. Correttezza, al seguente link puoi trovare alcuni esempi.

TUTTI i TESTI riportati sono in lettere MINUSCOLE, di conseguenza, **l'assenza di maiuscole non è da considerarsi un errore.**

In seguito alla valutazione, ti verrà chiesto di specificare una nota riguardo alla tua scelta. Un esempio di motivazione è **"La parola "parlargli" fa riferimento all'amico di cui si parla nel contesto del dialogo.".**

A: nelle note che hai scritto ieri leggo " relax e zero pensieri o quasi visto i ragazzi a casa, che comunque sono stati bravi.. . da rifare! " hai detto che ti sentivi felice per questo. adesso come ti senti?

B: triste, un solo giorno di ferie è troppo poco.

A: potresti provare a ricavarti dei momenti di relax anche in giorni lavorativi.

B: non saprei come fare, non ho molto tempo libero.

A: <continuazione proposta>

1 2 3 4

Appropriatezza

La seguente continuazione proposta è appropriata?

Identifichiamo quanto una continuazione è appropriata per un dato contesto oppure no, ossia se la continuazione è coerente e ha senso nel contesto corrente.

Risposta Passando con il mouse sopra ad ogni risposta farai apparire la possibilità che stai valutando nella conversazione originale alla tua sinistra	**Appropriata** La continuazione ha senso e può essere la naturale continuazione del contesto del dialogo mostrato.	**Non Appropriata** La continuazione non è coerente con il contesto del dialogo.	**Non so** Alcuni elementi sono appropriati altri meno.
A: mi spiace che tu ti senta così. prendi nota di come ti senti dopo averli passati.	○ Appropriata	○ Non Appropriata	○ Non so

PROSSIMA

Figure 3: Throughout the task, a short version of the guidelines is always presented to the annotator with the possibility to access the complete version via hyperlinks. During the evaluation, the corresponding dialogue context is shown to the annotator on the left, while the criterion question and the proposed response candidate are presented on the right, along with the name of the dimension, the definition of the dimension, and the possible decision values and their definitions. In order to reduce the cognitive workload of the annotators, all candidates for a specific dialogue context are evaluated one by one for the same criterion after one another (i.e. the annotator evaluates all the candidates of the presented dialogue history for criterion A, and then all the same candidates regarding criterion B). In this way, the left side of the UI (dialogue history) remains unchanged so that the annotator does not have to go through the dialogue history several times, and focuses on each evaluation metric per sets of response candidate.