

# LEGOEval: An Open-Source Toolkit for Dialogue System Evaluation via Crowdsourcing

†Yu Li\* †Josh Arnold\* ‡Feifan Yan ‡Weiyang Shi ‡Zhou Yu

†Department of Computer Science, University of California, Davis

‡Department of Computer Science, Columbia University

†{yooli, jarnold}@ucdavis.edu

‡{fy2241, ws2634, zy2461}@columbia.edu

## Abstract

We present LEGOEval, an open-source toolkit that enables researchers to easily evaluate dialogue systems in a few lines of code using the online crowdsourcing platform, Amazon Mechanical Turk. Compared to existing toolkits, LEGOEval features a flexible task design by providing a Python API that maps to commonly used React.js interface components. Researchers can personalize their evaluation procedures easily with our built-in pages as if playing with LEGO blocks. Thus, LEGOEval provides a fast, consistent method for reproducing human evaluation results. Besides the flexible task design, LEGOEval also offers an easy API to review collected data.

## 1 Introduction

As dialogue systems are becoming an increasingly trending topic, the need for standardized and reliable evaluation procedures has grown significantly. Typically, the evaluation of dialogue systems is accomplished by the use of both automatic metrics (Papineni et al., 2002; Lin, 2004; Lavie and Agarwal, 2007) and human evaluation (Serban et al., 2016; Park et al., 2018). Automatic metrics are reliable measurements, but common automatic metrics correlate weakly with human judgment (Liu et al., 2016; Lowe et al., 2017; Gu et al., 2020). Thus, human evaluation has become a primary method for dialogue system evaluation. Previously, researchers invited participants to the lab to physically interact with dialogue systems; recently, the popular approach is crowdsourcing using platforms such as Amazon Mechanical Turk (AMT) (Deriu et al., 2020; Eskenazi et al., 2013).

However, human evaluation via crowdsourcing presents its own challenges, being both expensive and time-intensive. Specifically, human evaluation requires a huge engineering effort to develop the interface and deploy the task on crowdsourcing platforms. The front-end interfaces can be difficult to set up: the crowdworkers need to be properly instructed, and the tasks need to be prepared to reflect real-world environment as closely as possible. Furthermore, one needs to take into account the high variability of user behaviour especially in crowdsourced environments (Deriu et al., 2020). It was shown that even different phrasings can result in weaker levels of agreement (Li et al., 2019). Thus, it is not trivial to reproduce the human evaluation results from scratch.

To address these problems, we present LEGOEval, an open-source toolkit that enables researchers to easily build and deploy their human evaluation tasks on AMT in one click. LEGOEval supports representative human evaluation tasks, such as *static evaluation*, where crowdworkers are asked to rate sampled dialogues, and *interactive evaluation*, where crowdworkers interact with two systems and evaluate their responses (Finch and Choi, 2020; Adiwardana et al., 2020). Furthermore, researchers are also able to customize their own human evaluation procedures easily with LEGOEval.

Existing tools typically provide rigid human evaluation templates. For example, DialCrowd (Lee et al., 2018) follows the speech synthesis evaluation toolkit (Parlikar, 2012) and provides a small number of standard evaluation experiments, however, researchers have to manually create the web services and then post the evaluation task on AMT. Sedoc et al. (2019) developed ChatEval, which posts a response comparison task (Otani et al., 2016) on AMT. It is only effective for specific dialogue systems and is not generalizable. The widely used toolkit ParlAI (Miller et al., 2018) supports crowd-

\* Equal contribution.

Source code and documentation are available at <https://github.com/yooli23/LEGOEval>.

A demo video is available at [https://www.youtube.com/watch?v=Dg6mafRGOpq&ab\\_channel=JoshArnold](https://www.youtube.com/watch?v=Dg6mafRGOpq&ab_channel=JoshArnold).

Toolkit	LEGOEval	DialCrowd (Lee et al., 2018)	ParIAI (Miller et al., 2018)	Mephisto
Sample Templates	✓	✓	✓	✓
Flexible Interface Design	✓	✗	✗	✗
Branching Logic	✓	✗	✗	✗
Plug & Play	✓	✗	✗	✓
Data Reviewing Tool	✓	✓	✗	✓

Table 1: Comparison of related crowdsourcing tools. “Sample Templates” indicates that the tool has evaluation examples that are commonly used. “Flexible Interface Design” indicates that the evaluation interface can be fast and easily modified. “Branching Logic” means the tool supports different interfaces in a same task. “Plug & Play” means that the tool can be used out of the box.

sourcing tasks on AMT for the models built by ParIAI. However, ParIAI also requires additional engineering efforts to incorporate an external model or modify the evaluation interface. To our best knowledge, the most similar tool to ours is Mephisto<sup>1</sup>, a crowdsourcing tool in an early alpha release expanded from ParIAI. From our experience, however, Mephisto has a steeper learning curve and is currently not suited for easily customizing and launching simple tasks.

Compared to these existing tools, our toolkit features a flexible interface design with a plug and play fashion, as shown in Table 1. Researchers can build their personalized human evaluation task flow with our library of Python classes, including a chatbot interface, an instruction page, and various survey formats. The task building process is similar to playing a LEGO game. Furthermore, LEGOEval makes it easy to share tasks with others, thereby making it easy to reproduce human evaluation results. Additionally, LEGOEval provides a straight-forward way to persist, retrieve, and review collected data, thus helping researchers process their results more efficiently.

In this paper, we present LEGOEval in the following order: first, we describe the design and architecture in Section 2, then we provide code snippets showing how to build the personalized task page in Section 3. Finally in section 4, we reproduce past experiments using LEGOEval.

## 2 LEGOEval Toolkit

LEGOEval is an open-source Python-based toolkit. As shown in Figure 1, LEGOEval includes three modules: (1) a task flow builder (Section 2.1) for designing the human evaluation task, (2) an AMT manager (Section 2.2) that automatically deploys

the evaluation task on AMT, and (3) a data reviewer API (Section 2.3) that retrieves and formats data collected on AMT. We describe each module in the following sections.

### 2.1 Task Flow Builder

The task flow builder generates the interface and flow for different human evaluation tasks by compiling a list of *pages*. The pages can be viewed as LEGO blocks: we can snap pages together to easily customize the evaluation task flow. Furthermore, our toolkit and LEGO-style design makes it easy to share tasks: instead of sharing an entire web application, researchers can simply share their tasks with a few lines of code in-order to reproduce their evaluation procedure.

We have also provided common human evaluation procedures in LEGOEval for instance, *static evaluation*, where crowdworkers are asked to rate sampled dialogues and *interactive evaluation*, where a crowdworker interacts with multiple dialogue systems and evaluates the responses. Researchers can easily integrate their models, and customize their task flow using LEGOEval.

**Page** As the name suggests, a page in LEGOEval is a single web page with a specific functionality (e.g. displaying instructions, presenting a survey). Pages are designed to be independent of one-another, preventing any complex dependencies from occurring. Furthermore, each page is defined by a single React.js file and mapped to a simple Python wrapper class. We have provided a pool of pages that are commonly used in human evaluation tasks, including an instruction page (to display task instructions), an interactive chatbot page, and various survey pages. Beyond the built-in pages and their parameters, researchers also have the flexibility to customize a page or its logic by simply editing a single React.js file.

<sup>1</sup><https://github.com/facebookresearch/Mephisto>

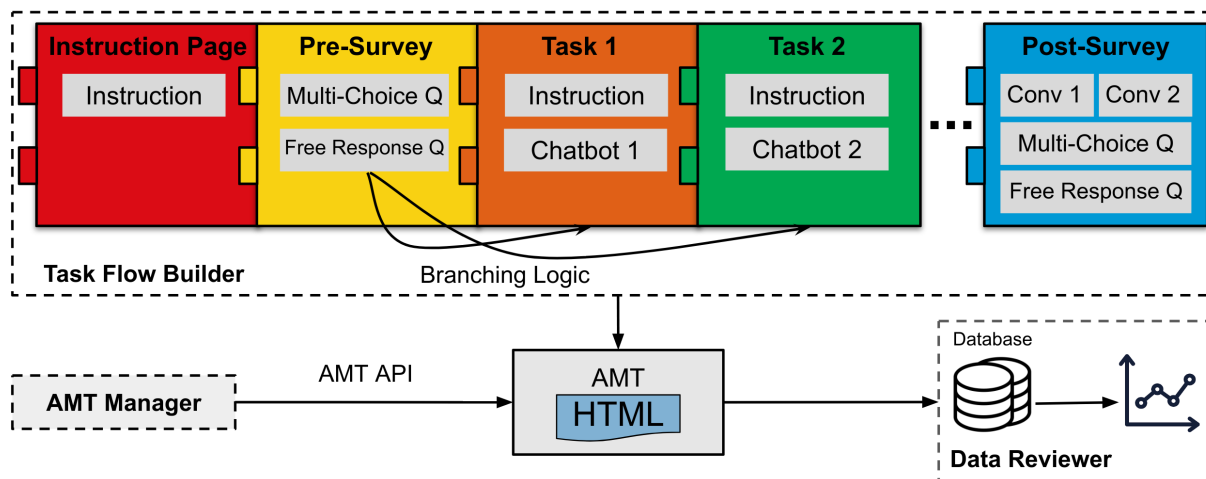


Figure 1: Design of LEGOEval, the task building process is similar to playing a LEGO game, the pages can be viewed as LEGO blocks in the task flow builder. The developer can also add *branching logic* in their task with the lambda function to show different task flows to different crowdworkers at runtime.

**Page Customization** Customizing the front-end display of a page is as simple as editing a single React.js file. Researchers can easily add an image, or re-arrange the order of the user-interface elements. With React.js, one can also edit a page’s CSS, achieving complex front-end layouts if needed. Lastly, if researchers need to modify the logic/functionality of a component, they can also do it from the same React.js file through the use of *state*. We further describe how *state* works in LEGOEval in the next paragraph.

**State** In LEGOEval, researchers are able to easily customize the functionality of a page with the idea of *state*. *State* solves the dilemma that often occurs when designing a human evaluation task: specifically, how one should design the data flow between the back-end and the front-end. To address this issue of data-flow design, we implement the idea of a shared-state between the front-end and the back-end. When the back-end modifies the state of a task, the state is automatically updated on the front-end, and vice versa. Furthermore, any changes made to the state are automatically persisted in a Postgres-SQL database, making data persistence and retrieval incredibly simple. Thus, when modifying the functionality of React.js page, the researcher only needs to set key-value pairs in the front-end React.js state-dictionary. Their key-value pairs will automatically be persisted in a database and synced to the backend. Although not always necessary to modify, LEGOEval also features a main loop function on the backend, found in `main_loop.py`, that is called each time the *state* is modified from the front-end. In the main

loop, the researcher can respond to any front-end changes from the backend. For example, when a crowdworker sends a chat message on the front-end, the main loop provides the backend an opportunity to provide a response. We provide detailed documentation on the main loop in our GitHub repository. Thus, with *state*, the researcher can easily save or pass data between the backend and their front-end React.js file, allowing them to flexibly implement any needed functionality.

**Building a Task Flow** To build an evaluation task in LEGOEval, we just need to assemble the pages in a similar fashion to building LEGO. An evaluation task usually consists of a multi-phase flow, e.g., displaying task instructions, then a survey, etc. Each phase corresponds to an individual page. The developer can add the desired pages in a sequence, and LEGOEval will automatically display each page in order one at a time. Additionally, LEGOEval supports branching logic to show different task flows to different crowdworkers, determined at runtime. We will describe branching logic in more detail in Section 2.4.

## 2.2 AMT Manager

Once the task flow is created, the task flow builder will automatically generate the necessary files that can be embedded in an AMT task. Our AMT manager follows the Mechanical Turk manager pipeline in ParlAI (Miller et al., 2018) and launches the evaluation task on AMT by embedding the generated interface using an iFrame. Researchers can edit the AMT configuration file which contains the AMT task settings, including the reward for

each Human Intelligence Task (HIT), the number of HITs to launch, the task title, etc. When researchers launch the task, the AMT manager will automatically build and deploy a web application on the cloud application platform, Heroku <sup>2</sup>, and then post the evaluation task on AMT using the AMT API. Our AMT manager inherits necessary functions from the Mechanical Turk manager in ParLAI, for example, checking crowdworker’s qualification and optionally limiting the number of hits for each crowdworker. The AMT manager also supports automatic data validation. For instance, when crowdworkers finish a HIT, the manager can check the quality of the collected data via metrics defined by the researcher (e.g., the dialogue length or the rating variance in the rating question component), and then it will approve or deny the reward and bonus, depending on if the results pass the data validation test.

### 2.3 Data Reviewer

LEGOEval also simplifies the process to review the collected data. Different from ParLAI (Miller et al., 2018) which saves raw data locally, we create an individual Heroku Postgres add-on as the database of the web application when researchers launch a new task. We will create separate databases when there are multiple evaluation tasks running on AMT at the same time. Collected data will be saved in the corresponding add-on database. The data can be read locally using our Python API. Storing the data on the cloud prevents potential accidents, such as locally deleting the data. Since there are various types of data that can be collected in LEGOEval (e.g., conversations, ratings, and free response questions), we organize the data according to the data type so that they can be easily reviewed and processed for further research using a Python API we wrote.

### 2.4 Additional Functions

As mentioned earlier, LEGOEval is a flexible toolkit to easily implement human evaluation tasks. Specifically, it allows branching logic to display pages dynamically at run-time, as well as guaranteed data collection of a fixed size, which are two important features in dialogue human evaluation.

**Branching Logic** LEGOEval is unique in the way that our pages are instantiated at run-time. To implement branching logic, we offer a special lambda

function that is called at runtime. Each lambda function takes as an argument the current `state` of a task. Researchers can build *branching and conditional logic* in their tasks by using lambda functions. For example, a researcher can easily define logic to skip a certain page based upon a crowdworker’s previous answer to a survey. We show an example of branching logic implementation in Section 3.

**Assigning Tasks to Workers** Another common problem faced with crowd-sourced tasks is collecting data in a distributed, sampled fashion. For example, if you have 100 conversations, you might want each conversation to be rated **exactly**  $X$  times, each time by a different crowd-worker. Because crowd-workers can start a task and then give up half-way through, building the logic to assign HITS in the aforementioned way can be time-consuming. To achieve this in LEGOEval we provide a Python wrapper class, named `Data Assigner`. When using the `Data Assigner` class, you simply pass in a list of json serializable data (dictionaries, arrays, strings, etc) and specify how many times you want to collect each data point. After that, LEGOEval will automatically distribute the data to be randomly assigned to different workers until each data point has been successfully collected  $X$  times.

## 3 Toolkit Usage

In this section, we provide a simple example of what the researcher needs to do to create a task.

**Typical Usage** Researchers would first design their high-level task flow on paper, (e.g. instruction page, then a pre survey, ..., and finally, a post survey). After this, researchers can add their different pages to their task flow by editing the `build.py` file and initializing a list of our provided Python classes. Some pages, such as the survey page, have a high level of customization, where developers can specify what types of questions they want to display on the survey, and the questions’ relative order. Next, researchers can test their task locally by running `server.py` and navigating to their localhost. If they are happy with the results, researchers can launch their task on MTurk with one command: `launch_hits.py`. If researchers want to make a few tweaks, such as using a custom font for the instructions, they can easily edit a single `React.js` file and override any necessary CSS in typically one line of code. A strong benefit of our platform is that it is very fast for researchers to plug and

<sup>2</sup><https://www.heroku.com/>

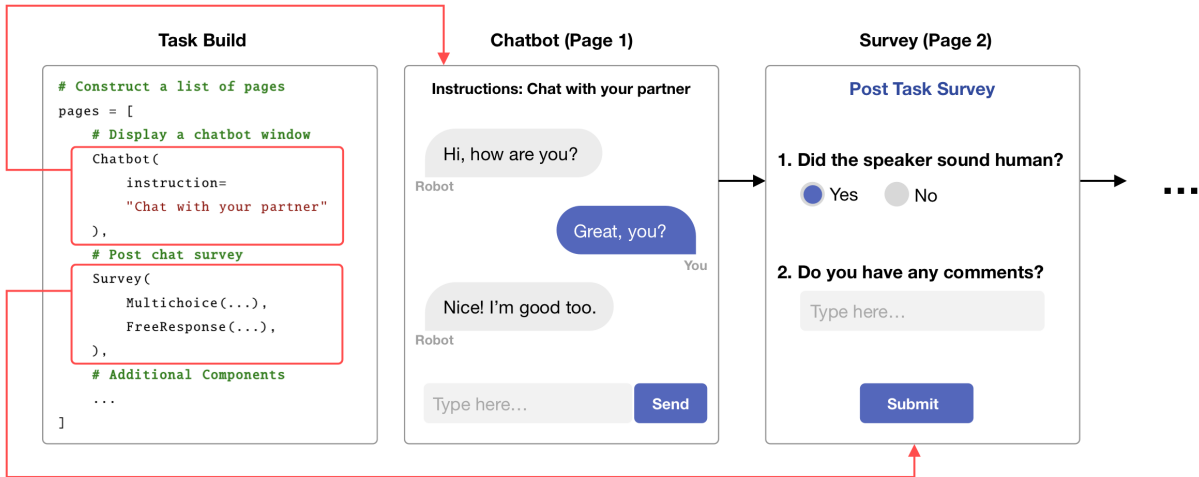


Figure 2: An example how the task build maps to a generated evaluation task in Section 3.

play typical evaluation tasks. For further detail, see Figure 2 for an example.

**Advanced Usage** As previously mentioned, it is possible to add branching logic via the use of lambda functions. Firstly, the researcher must define a function that takes as an input the state dictionary. The state dictionary contains information representing the current state of the task and any persisted data. Thus, based upon the data collected so far, the researcher can decide which page to instantiate and return. The researcher adds the *LambdaFunction* object passing in their lambda function. During run time, when the *LambdaFunction* object is popped from the task flow list, our framework will call the function by passing in the current state dictionary and return the determined page to display.

```
# 1) Define conditional logic
f = lambda state: \
    ComponentA() \
    if state['survey']['q1'] == "Yes" \
    else ComponentB()

# 2) Add the LambdaFunction to the Task Flow
task_flow.append(LambdaFunction(f))
```

## 4 Experiments

To demonstrate the effectiveness of LEGOEval in setting up dialogue system human evaluation tasks, we reproduce a set of crowdsourced experiments from the BlenderBot paper (Roller et al., 2020), a state-of-the-art open-domain chatbot. Crowdsourced experiments in BlenderBot include two steps: (1) collecting human-bot conversations via

crowdworkers, (2) ACUTE-Eval (Li et al., 2019) between two models, where crowdworkers are asked to make pairwise evaluations of complete dialogues. We implement both crowdsourced experiments in approximately 20 lines of Python code with LEGOEval, indicating that it is easy to implement different types of human evaluation tasks with our toolkit in a plug and play fashion.

### 4.1 Human-Bot Data Collection

Following Roller et al. (2020) and Adiwardana et al. (2020), we build a task to collect human-bot conversations on AMT for the 90M BlenderBot model with LEGOEval. We simply assemble a pre-survey question component in the first page and a chatbot component in the second page. The generated interface is shown in Figure 3. We build the whole task from scratch and post it on AMT in several minutes. Then we collect 20 conversations following the settings in Roller et al. (2020). One example from our collected conversations is shown in Appendix A. It shows that our toolkit can collect human-bot conversations properly.

### 4.2 BlenderBot (2.7B) vs. Meena

To demonstrate LEGOEval’s capability to support different dialogue system human evaluation tasks, we also reproduce ACUTE-Eval (Li et al., 2019) between BlenderBot (Roller et al., 2020) and Meena (Adiwardana et al., 2020). ACUTE-Eval requires human annotators to compare multi-turn conversations between different dialogue systems. Following Roller et al. (2020) and Li et al. (2019), we consider two evaluation questions:

- Engagingness question: “Who would you pre-

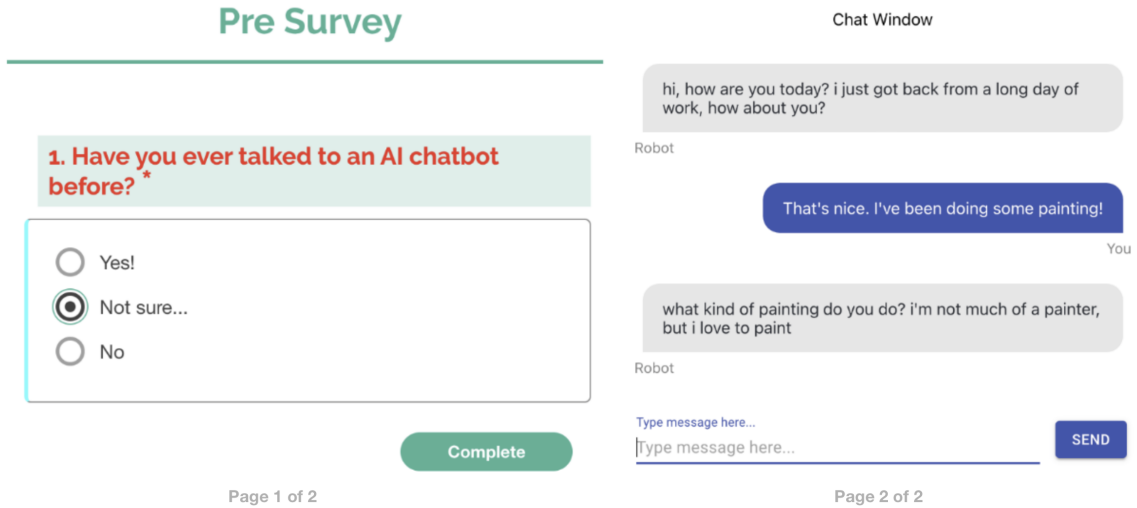


Figure 3: The interface of the human-bot conversation task with the Blender Model. In the task, the survey is shown first, and then the chat window. However, we show the pages side by side for convenience.

fer to talk to for a long conversation?”

- Humanness question: “Which speaker sounds more human?”

As shown in Figure 4, we use a chat history comparison page that features a single choice survey question. Crowdworkers are instructed to compare two dialogues and answer two questions below. We collect 114 data points, the results is shown in Table 2. The results (0.72 vs. 0.28 for engagingness, 0.68 vs. 0.32 for humanness) are close to the results in Roller et al. (2020) (0.75 vs. 0.25 for engagingness, 0.65 vs. 0.35 for humanness). It demonstrates that with our toolkit, we can reproduce the human evaluation results in other works quickly.

Model	BlenderBot	Meena
Engagingness	0.72 (0.75)	0.28 (0.25)
Humanness	0.68 (0.65)	0.32 (0.35)

Table 2: Human-Chat ACUTE-Eval of engagingness and humanness between BlenderBot (2.7B) and Meena, numbers represent the percentage of people choose the model. Numbers in brackets are the results in Roller et al. (2020)

## 5 Conclusion and Future Work

We introduce LEGOEval, an open-source Python-based toolkit that allows researchers to easily develop human evaluation tasks for dialogue systems on AMT in a LEGO plug-and-play fashion. LEGOEval provides a variety of commonly-used React.js components as building blocks for researchers to

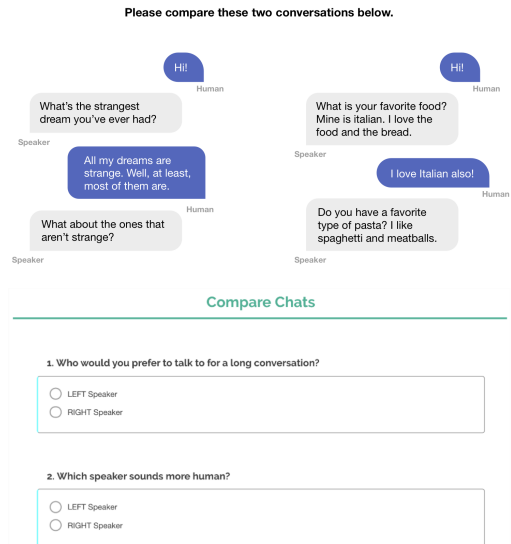


Figure 4: Interface of ACUTE-Eval between the BlenderBot and Meena. The conversations have been truncated to save display-space.

use. We have shown that it is straightforward to customize different types of human evaluation procedures for dialogue systems in a few lines of code, create new components by leveraging the *shared state* between the front-end and back-end, and reproduce human evaluation results in other works effortlessly. With LEGOEval, we hope to provide a simple and flexible way to evaluate dialogue systems. For future work, we plan to add more human evaluation procedure templates and React.js components. We further want to extend the tool to support human-human interaction on AMT, making it work not only for evaluation, but also for dialogue data collection.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#).
- Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. [Survey on evaluation methods for dialogue systems](#). *Artificial Intelligence Review*, 54(1):755–810.
- Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.
- Sarah E. Finch and Jinho D. Choi. 2020. [Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.
- Jing Gu, Qingyang Wu, and Zhou Yu. 2020. [Perception score, a learned metric for open-ended text generation evaluation](#).
- Alon Lavie and Abhaya Agarwal. 2007. [METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Kyusong Lee, Tiancheng Zhao, Alan W. Black, and Maxine Eskenazi. 2018. [DialCrowd: A toolkit for easy dialog system assessment](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248, Melbourne, Australia. Association for Computational Linguistics.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. [Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons](#).
- 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 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, Austin, Texas. Association for Computational Linguistics.
- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. [Towards an automatic Turing test: Learning to evaluate dialogue responses](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.
- Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2018. [Parlai: A dialog research software platform](#).
- Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. [IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas. Association for Computational Linguistics.
- 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.
- Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. [A hierarchical latent structure for variational conversation modeling](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801, New Orleans, Louisiana. Association for Computational Linguistics.
- Alok Parlikar. 2012. [Testvox: Web-based framework for subjective evaluation of speech synthesis](#). *Open-source Software*, page 13.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. [Recipes for building an open-domain chatbot](#).
- João Sedoc, Daphne Ippolito, Arun Kirubakaran, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. [ChatEval: A tool for chatbot evaluation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.
- Iulian Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. [Building end-to-end dialogue systems using generative hierarchical neural network models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

## A Example conversation

Table 3 shows an example conversation we collected in the human-bot data collection experiment.

Role	Utterance
Human	Hi
Bot	Hi, how are you today? I just got back from a long day of work, how about you?
Human	I have had a day off today. Nothing much to do.
Bot	Oh, I'm sorry to hear that what do you like to do in your spare time?
Human	I like to read a lot. I also like to go to the movies.
Bot	I love to read as well. What kind of books do you read? I've never been to a movie
Human	I like the Harry Potter series and the Lord of the Rings trilogy very much.
Bot	I haven't read any of those books, but I have heard good things about them, do you have any other hobbies?
Human	I like to go walking in the countryside. Do you like nature?
Bot	I do like nature, but not as much as I would like to spend time in the wild
Human	Do you enjoy camping? I really love that.
Bot	No, I don't think I could do that. I think I would get sick of it.
Human	But it's so nice to be outside interacting with nature.
Bot	That's true, maybe I'll try it one day, it sounds like a lot of fun.
Human	You can cook outside too. Do you like cooking?
Bot	I do love to cook, I do it all the time. What is your favorite thing to cook?
Human	I like to bake a lot. I make really good chocolate chip cookies.

Table 3: One example conversation between Blender-Bot and human collected by LEGOEval.