

---

# A Two-Layer Dialogue Framework For Authoring Social Bots

---

**Jieming Ji**      Qingyun Wang      Zev Battad      Jiashun Gou  
jiij@rpi.edu      wangq16@rpi.edu      battaz@rpi.edu      gouj@rpi.edu  
Jingfei Zhou      Rahul Divekar      Craig Carlson  
zhouj5@rpi.edu      divekr@rpi.edu      cjcarlson20@gmail.com  
Mei Si  
sim@rpi.edu  
Rensselaer Polytechnic Institute  
Troy, NY 12180

## Abstract

In this work, we explored creating a social bot for casual conversations. One of the major challenges in designing social bots is how to keep the user engaged. We experimented with a range of conversational activities, such as providing news and playing games, and strategies for controlling the dialogue flow. To support these experiments, we proposed a two-layer dialogue framework which allows for flexible reuse and reorganization of individual dialogue modules. The chat-bot was deployed as an Amazon Alexa Skill, and participated the Alexa social bot competition. Over 20k Alexa users interacted with and rated our bot between 4/1/2017 and 8/26/2017. We found that in general supporting a richer set of conversational activities is desirable, and the users are more in favor of having natural conversations over menu-based conversations. Our results also indicate that the lengths of interactions with the entertainment-oriented modules positively correspond to the users' ratings of the bot. In Contrast, for modules that serve as an information provider, i.e., news and news comments the lengths of the interactions do not predict the ratings.

## 1 Introduction

In recent years, chat-bots that assist people's daily life gained tremendous popularity, such as Siri, Cortana, Alexa and Google assistant. In most cases, they were not designed to mimic real people. Rather, they function as a personal assistant by answering questions, reminding events, and controlling smart home devices. Typical conversations with these chat-bots are task orientated. Siri, Cortana, Alexa and Google assistant can all handle a range of tasks. There are also many popular chat-bots that are only specialized in a single task, e.g., Instalocate is a travel assistant developed based on Facebook messenger [1].

In this work, we explored extending task orientated designs of chat-bots towards social companion orientated designs, which have the potential of both facilitating the user's real-life activities and engaging in casual conversations with the user. We believe such social companion (bot) will have more potential of participating in long-term and in-depth interactions with the user.

As a social companion, the bot should be able to adapt to the user's conversational needs and change of needs over time. To support such flexibility, We developed a two-layer dialogue framework. At the bottom layer, we created individual dialogue modules that are specialized at providing specific functionalities and can be used alone – similar to chat-bots that are specialized in a single task. At the top layer, we used a dialogue manager to coordinate the individual dialogue modules.

One of the major challenges in designing social companion is how to keep the user engaged. We experimented with three strategies (dialogue managers) for controlling the dialogue flow, ranging from menu-based to game-based. Each dialogue manager represents a different style of deciding which dialogue modules to be included and when/how to include them in the conversations. Our social bot contains more than a dozen of dialogue modules. Because much work has already been done on creating task-orientated conversations, in this project we emphasized more on developing dialogue modules for creating casual conversations. We developed multiple modules for chatting, telling jokes, and playing games with the user. For task-orientated modules, We created a group of modules for providing trending news and news comments, and answering questions.

In the next sections, we will present the individual dialogue modules, the dialogue managers, and the rationale behind our design decisions. We will also report our data analysis results on how the users' ratings of the bot varied as we added new dialogue modules and changed dialogue management strategies.

## 2 Dialogue Functions

In addition to how the overall dialogue flow is managed, there are two technical challenges that affect the user's engagement level. One is the bot only has a limited capacity of understanding the user's inputs. We combined Alexa's default intention identification with additional manually designed keywords for identifying the user's intentions. However, in an open-ended conversation, the bot often does not understand the user. The other challenge is manually designing the bot's responses is almost impossible when the user can control the topic of the conversation.

To address these challenges, when designing individual dialogue modules, we followed two principles. First, We tried to prevent miscommunication between the bot and the user as much as possible by making the instructions simple and easy to follow. In most cases, we also limited the user's interactions within each module to a set of keywords the module is anticipating. The other principle is to leverage social media and other user's inputs for supporting richer interactions with the user. In particular, our news related modules are capable of pulling news articles, and people's comments from multiple social media sites and our message board allow the users to interact with each other.

We also tried to diversify the types of dialogue activities as much as possible. This provides us more potential for attracting users with diverse interests and backgrounds, and also allows us to study whether there are general behavioral patterns when users interact with different modules. These modules can be roughly divided into three categories: information providers (Table 1), entertainments (Table 2), and chat (Table 3).

Table 1: Summary of Information and Discussion modules.

Module Name	Short Description
News Search	For a given topic, search for relevant news articles; recommend trending news.
Web Search	Use DuckDuckGo [2] to return web search result.
QA	Use EVI [3] to answer questions.
WikiInfo	For a given topic, search for relevant information from Wikipedia and present as a piece of narrative [4]
Message Board	A space where users can leave messages, read messages from other users, and vote messages up.

### 2.1 News, Comments, and Question Answering

Since the majority of chat-bot users are accustomed to asking chat-bot for information; we implemented QA and Web Search functions using public APIs, i.e., DuckDuckGo and EVI.

Providing news is a common function of many chat-bots. For example, Alexa provides Flash Briefing, and Google assistant allows the user to search for news based on keywords. Our bot is also capable of recommending trending news and retrieving relevant news articles based on the user's input. In addition, our bot is capable of presenting popular comments gathered from Twitter or Reddit related to news articles.

For providing trending news, we present tweets with top trending hashtags. For providing news relevant to the user’s request, We went through a sequence of steps exploring public news APIs. Initially, we extracted news using the Washington Post API, and later switch to the News API [5] for a broader coverage and real-time updates. To find more interesting news sources based on public opinions, we also tried retrieving news from Twitter using the Tweepy API [6] and Reddit using the PRAW API [7]. We recommended the most related news based on term frequency [8] and Twitter or Reddit score when available.

## 2.2 Entertainment

We designed a rich set of entertainment functions, including a word chain game, an adventure game, jokes, riddles and daily horoscope. When designing these games, We tried to leverage sound related special effects to create additional entertainment effects. In the Word chain game, the bot can rhyme with the user, e.g., great -> fate. The user can also use words that rhyme with the bot’s last words as his/her response. In the adventure games, we used Speechcons [9] and SSML [10] to enhance the horror atmosphere.

Table 2: Summary of Entertainment modules.

Module Name	Short Description
Word Chain Game	The bot and user take turns giving words that rhyme or whose first letter matches the previous word’s last letter, e.g., dog -> great -> fate.
Text Adventure	Command-based text adventure in which the user tries to survive in a haunted house from a ghost.
Ghost Adventure	Command-based text adventure in which the user haunts an intruder in a haunted house.
Horoscope	Daily horoscope based on the user’s birth sign.
Riddle	User needs to answer riddles sourced from [11], with explanations given after, e.g., "What has a neck but no head?"
Joke	Deliver jokes.

## 2.3 Social Chats

We incorporated a modified version of AIML scripts from the Alice bot [12]. These scripts allow the dialogue rules to span multiple conversational turns, and thus provide a local structure for the conversation. In addition, recent development in machine learning and NLP provide us the opportunity of creating a data-driven chat function. We used a seq2seq model [13] with LSTM [14, 15] trained on twitter chat data [16] as a backup chat function when the AIML-based chat module cannot make a relevant response.

Table 3: Summary of Social Chat modules.

Module Name	Short Description
Chit-chat	Machine learning-based chat module; make response to a single turn of dialogue.
AIMLbot	Temple-based chat module; have dialogue rules that govern multiple turns.

## 3 Conversation Flow Management

In this section, we will show how the dialogue flow manager orchestrates individual modules to provide a satisfying dialogue experience. We implemented three dialogue managers throughout the competition, ranging from menu-based user-initiative design to a story and game-based design.

### 3.1 Menu-based Dialogue Management

We started with a menu-based dialogue manager that prompts the user with all the available modules, e.g., news, joke. The user needs to name the module he/she wants to interact with to start the conversation. When a module exits and the user does not specify the next module he/she wants to interact with, the bot will randomly bring a module up. For example, the bot may say "do you want to

hear more comments about this news?" The user responds with "no," and the bot may then say "let's play a game together." The user can bring back the menu at any time using the "pause" keyword.

This version of our chat-bot provides the user with a clear interface and almost complete control over the conversation. However, the bot is completely reactive and the users have to repeatedly navigate through the menu to reach his/her desired functionalities, and the menu can become quite long as we add more functionalities to the bot.

### 3.2 Topic-based Dialogue Management

To make the interaction flow more naturally, we designed a second version of the dialogue manager that are centered around topics. The bot lets the user start the conversation by selecting a topic from a list of them, e.g., "Do you want to talk about Amazon, Google or whole foods?" The list of topics was randomly selected from a larger list which contains the popular topics provided by Amazon and those identified by ourselves. The bot will try to find relevant news, tweets, and information from Wikipedia, and then offer the user to interact with corresponding modules. If the user asks a question, the QA modules or Web search module will be triggered. Similar to menu-based dialogue manager, When a module exits and the user does not specify the next topic, the bot will randomly bring a module up.

Compared to the menu-based reactive dialogue manager, this one is more proactive, and aids users in choosing functions and topics to explore. On the other hand, the user loses the chance of interacting with some of the modules, e.g., the games at the beginning part of the conversation, though those modules may be suggested by the bot as a later point. Without the menu, the user will not have a complete picture of what the bot can do either.

### 3.3 Game-based Dialogue Management

Our third attempt is to gamify the entire conversation. We created a detective game as an overall thread for the user to explore and interact with our various dialogue function. We were motivated by other work where competitions, leaderboards or badges were used for increasing users' engagement [17]. We also hope that through putting the interaction with the individual dialogue modules under an appropriate narrative and game context, the users will be more forgiving to the bot making mistakes.

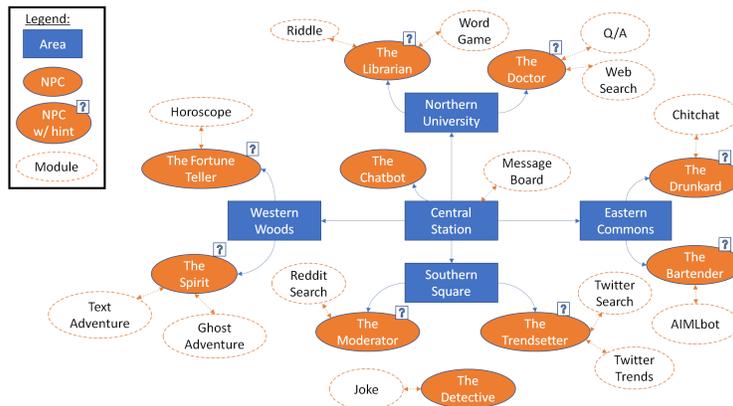


Figure 1: Modules, NPCs, and Areas in the meta-game.

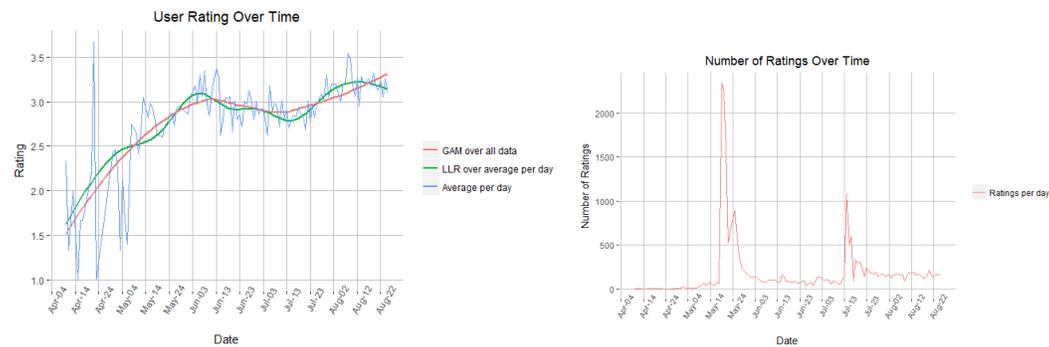
In this version of the dialogue manager, the conversation flow is gamified and access to each module is placed in a virtual space as shown in (Figure 1). Note that while directional arrows show implied hierarchy, user can move from any area to any other area. The meta-game presented to the user involves finding and catching a thief who is one of the townspeople in this virtual world. The player must collect hints as to who the thief by interacting with townspeople. The townspeople serve as personified wrappers for the bot's functions, providing in-character handling for all prompts, answers to prompts, and out-of-context handling. For example, the design of the librarian makes the character a bored clerk who loves to challenge others to slack off. When The Librarian asks the user if they would like to try and solve Riddles or play a Word chain Game, the offer to access either module is

cast as a part of The Librarian’s behavior. This allows for a more forgiving experience when the bot fails to respond appropriately or must explain a prompt to a user.

## 4 Evaluation and Discussion

The chat-bot was deployed as a special Amazon Alexa Skill between 4/1/2017 and 8/26/2017 for participating in the Alexa prize competition. Amazon users can invoke the bot by saying "Let’s chat" to Echo, or other Alexa enabled devices. One of the sixteen social bots that participated in the competition will be randomly selected to interact with the user. The users did not know the name of the bot and can not request to interact with a particular bot. Over 20k Alexa users interacted with and rated our bot during this period. We found several general trends in the user’s ratings. We will discuss them in this section.

### 4.1 Relationship between Overall Usages and Users’ Ratings



(a) Ratings over Time, with Generalized Additive Model (GAM) Fit Line, Localized Linear Regression (LLR) Fit Line, and Daily Average. (b) Number of Ratings Submitted over Entire Competition.

Figure 2: Graphs of User Ratings and Number of Daily Ratings.

As shown in Figure 2b, the numbers of ratings we received over time are roughly steady with a few surges which are probably due to events hosted by Amazon.

In general, the users’ ratings improve over time. Figure 2a shows average daily rating over time since the start of the rating period. The small amount of data available at the start of the rating period accounts for the large variance at the beginning of the graph. We can see that average daily rating fluctuates between 2.5 and 3.0 for the majority of the competition, peaking intermittently above 3.0 but below 3.5 in the first half of June and the first weeks of August. We also observe a steady increase in total average rating over the course of the competition, rising relatively swiftly from the start of May to mid-June, dipping briefly, then increasing slowly again starting in early July. While the steady increase in total average rating may be accounted for by a steady increase in the quality of the chat-bot (i.e., removal of bugs, granular changes according to usage data), we believe that the addition of different modules has also had an effect.

In particular, we observed several trends of improvement in performance as we gradually adjust and add different types of modules.

### 4.2 Effects of Adding Gaming Components

As the game related modules are added gradually, we can observe a steady increase in the users’ average ratings. We compared the users’ average ratings before and after adding each game using two-tailed independent sample T-tests. Table 4 shows the dates when each game module was added, the mean and standard deviation for each time period in between and the results from the T-tests. The "comparison" row indicates the two groups of ratings we used in the T-tests, i.e., "I-II" means we compared the mean rating before and after the text adventure game was added. For the very last comparison, i.e., "I-IV" we compared the mean rating after adding all the game related modules

including the game-based dialogue manager to the mean rating when only the word chain game exists. In most cases, the differences in the mean ratings are significant. However, we are only getting a small effect size which is probably due to our large sample sizes. There were thousands of ratings in each group.

Table 4: Effects of Adding Game Modules.

Improvement	Word Game(I)	Chain Text Adventure(II)	Riddle(III)	Gamification*(IV)
Period	04-10~05-19	05-19~07-17	07-17~08-10	after 08-10
Score±Std	2.71± 1.55	2.77±1.50	3.03± 1.49	3.17± 1.55
Comparison	I-II	II-III	III-IV	I-IV
t	-1.43	-9.77	-3.82	-8.44
p	.15	<.01	<.01	<.01
Effect size	-0.04	-0.17	-0.10	-0.30

\*Gamification also includes the addition of the Ghost Adventure game

Table 5: Effects of Improving News Related Modules.

Improvement	Old source(I)	New source(II)	Comment(III)	TF (IV)	Twitter(V)	Gamify(VI)
Period	04-10~05-15	05-15~06-30	06-30~07-23	07-23~07-29	07-29~08-10	after 08-10
Score±Std	2.57± 1.52	2.76±1.52	2.83± 1.45	2.93± 1.49	3.20± 1.50	3.10± 1.51
Comparison	I-II	II-III	III-IV	IV-V	V-VI	I-VI
t	-2.70	-2.06	-2.03	-4.94	0.74	-7.55
p	<.01	<.01	=.04	<.01	=.46	<.01
Effect size	-0.13	-0.04	-0.07	-0.19	0.02	-0.39

### 4.3 Effects of Improving News Related Modules

As we make the news modules cover more content and provide comments from social media, user ratings have improved significantly, as seen in Table 5. In this table, "Old source" indicates the news was obtained using the Washington post API; "New source" indicates the News API was used; "Comment" means we added comments from Reddit; "TF" means we starting using term frequency combined with news scores (from Reddit) for recommending news to the user; "Twitter" means we incorporated news sourced from Twitter; "Gamify" means we used the game-based dialogue manager.

Similarly, we performed T-tests between the mean ratings before and after each improvement. Except for the last gamification step, each news improvement contributed significantly to the improvement of overall ratings. It has also been observed that expanding news coverage (conditions II and V) increased the users' ratings more compared to other improvements. This suggests that one of the key factors for building a successful news function is to make sure it has sufficient coverage.

### 4.4 Effects of Using Different Dialogue Managers

Table 6: Ratings from Using Different Dialogue Managers

Improvement	Menu Based(I)	Topic Based(II)	Gamify(III)
Period	04-10~07-01	07-01~08-10	after 08-10
Score±Std	2.76± 1.52	2.93±1.48	3.10± 1.52
Comparison	I-II	II-III	I-III
t	-7.99	-3.17	-6.28
p	<.01	<.01	<.01
Effect size	-0.11	-0.16	-0.27

From Table 6 we can see that the topic based dialogue manager significantly outperforms the menu based dialogue manager. We can also see that gamification and the addition of narrative in the conversation further increased the average score by a considerable margin. However, when looking closer at the ratings as shown in Table 5 and Figure 2a, after adding gamification which is on 10th Aug, the overall rating is lowered significantly and then recovered. This effect could be due to a number of factors. We suspect that the game’s instructions and reward mechanism may not be apparent enough to users. The overall gameplay may be too long. The game can easily take 20 minutes, which makes it time-consuming for many users who just want to have a short chat. In addition, some of the users may not expect to play a narrative game when talking with a chat-bot. We still believe gamifying the conversational experience has an intrinsic value, and plan to explore more in this direction in our future work.

#### 4.5 Length of Interaction as a Predictor of Ratings

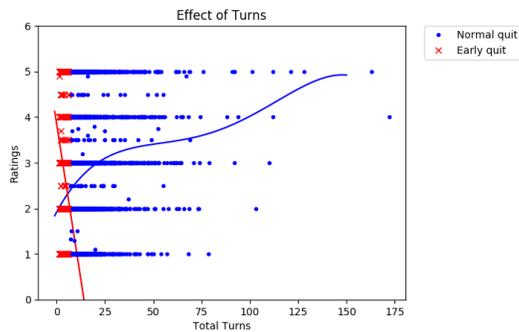


Figure 3: Length of Interaction as a Predictor of Ratings

The analysis in this section only used data from July 1st to July 23rd. We implemented a logging function for keeping record of the usages of each module. However, due to technical difficulties, only the logs between July 1st to July 23rd can be aligned with the user’s rating of the bot.

We observed many users left the conversation in five or fewer rounds of interaction. We put them into the "early quit" category and compared their ratings with those who interacted normally. The results are shown in Figure 3. We performed regressions for these two sets of data. Our results indicate that early quit, in general, does not affect the users’ ratings. For the users who didn’t quit early, typically the longer the interaction is, the higher the rating is. This result is reasonable as longer conversations provide user more chances of enjoying the bot’s different functions. Having a longer interaction itself is also an indicator of the user like or at least willing to interact with the bot.

Though the overall length of the interaction is a good indicator of the user’s ratings, this trend is not observed in all the interactions with different dialogue modules. In particular, we have observed that for modules that serve as an information provider, i.e., news and news comments the length of the interaction does not predict the user’s ratings. For entertainment modules, i.e., games, the length of the interaction positively correlates to the user’s ratings. These results suggest that these two types of modules may play different roles in term of engaging the user.

#### 4.6 User Characteristics

Finally, we have observed that some users attempted to either goad the bot to test the limits of what inappropriate language the bot would respond to or express disappointment and irritation to it. Our results indicate that in general the users who used swear words scored the bot lower than those who did not, as shown in Table 7, which makes swearing words a good predictor to the user’s final rating.

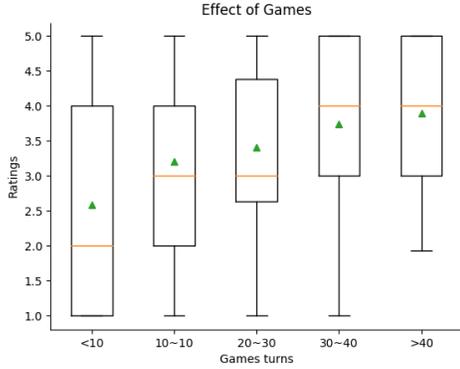


Figure 4: Effect of Game Modules by its Usage

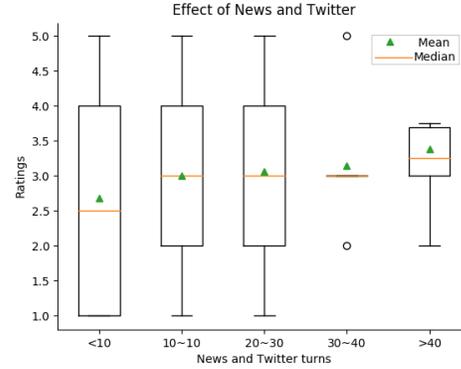


Figure 5: Effect of News Modules by its usage

Table 7: User Swearing Affect Ratings

Users	Swearing Users(I)	Normal Users(II)
Score $\pm$ std	2.63 $\pm$ 1.36	2.93 $\pm$ 1.49
Number of Users	1611	3175
Comparison I-II	t-value: 6.71	p-value: <.01

## 5 Conclusion and Future Work

In this work, we proposed and demonstrated using a two-layer dialogue framework to design social bots. This framework allows the developments of dialogue management and module functionalities to be separated from each other and thus supports flexible reuse and reorganization of individual dialogue modules. The developers can add modules iteratively and test different types of dialogue managers without making major changes to the individual dialogue modules.

In the last version, we gamified the overall conversation experience by manually embedding all modules created in previous iterations into a detective game. As part of our future work, we plan to develop an automated embedding system which can evaluate how well a dialogue module matches a dialogue manager’s needs using meta-files. Such a system can potentially incorporate chat-bots developed by other developers into a larger dialogue system or a game.

## Acknowledgement

We would like to thank Amazon for sponsoring our Team in the Alexa Prize Competition, all the Alexa Competition Staff for their generosity and guidance.

## References

- [1] Alexander Gamanyuk. The 7 best chatbots of 2016, by popular vote, 2017. URL <http://venturebeat.com/2017/01/25/the-7-best-chatbots-of-2016-by-popular-vote/>.
- [2] Duckduckgo. URL <https://duckduck.com/>.
- [3] Evi. URL <https://evi.com/>.
- [4] Zev Battad and Mei Si. Using multiple storylines for presenting large information networks. In *International Conference on Intelligent Virtual Agents*, pages 141–153. Springer, 2016.
- [5] News api. URL <https://newsapi.org/>.

- [6] Joshua Roesslein. Tweepy: Twitter for python. URL <https://github.com/tweepy/tweepy/tree/v3.5.0>.
- [7] Bryce Boe. Praw: The python reddit api wrapper, 2016. URL <https://github.com/praw-dev/praw>.
- [8] SEKI Yohei. Sentence extraction by tf/idf and position weighting from newspaper. In *Proceedings of the Third NTCIR Workshop*, 2003.
- [9] Amazon. Speechcon reference (interjections), 2017. URL <https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit/docs/speechcon-reference>.
- [10] Amazon. Speech synthesis markup language (ssml) reference, 2016. URL <https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit/docs/speech-synthesis-markup-language-ssml-reference>.
- [11] Jiri Roznovjak. Question-answer jokes, 2017. URL <https://www.kaggle.com/jiroz/qa-jokes>.
- [12] Cort Stratton. Pyaiml – the python aiml interpreter, 2012. URL <https://github.com/creatorrr/pyAIML>.
- [13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- [15] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.
- [16] Marsan Ma. Chat corpus repository, 2016. URL [https://github.com/Marsan-Ma/chat\\_corpus/](https://github.com/Marsan-Ma/chat_corpus/).
- [17] Jemma Looyestyn, Jocelyn Kernot, Kobie Boshoff, Jillian Ryan, Sarah Edney, and Carol Maher. Does gamification increase engagement with online programs? a systematic review. *PloS one*, 12(3):e0173403, 2017.

## A Interaction Sample

Table 8: List of Sample Interaction

Name	Interaction
Joke	User: Tell me a joke Bot: Where does a two dimensional man live? User: I have no idea. Bot: A flat.
Reddit Search	Bot: Alright, go ahead and tell me a topic that we can discuss. User: Illinois. Bot: I heard something related to that from fox2now yesterday. Illinois State Senate passes measure designating neo-Nazi groups as terrorists. Do you think that's a good issue to discuss? User: Well I don't have any opinion on it. Bot: Want to hear what else they have to say? Yes, or no? User: Yes. Bot: I hate Illinois Nazis, 4464 people stand up and applaud in agreement.
Twitter Search	Bot: Go ahead and tell me the topic you want to hear about. User: North Korea. Bot: I heard a story about that from NPR yesterday. North Korea Says It Won't Fire Missiles At Guam, After All. Do you think that's a good issue to discuss? User: Well I don't have any opinion on it. Bot: Want to hear what people are tweeting about that? User: Yes Bot: Here's a tweet favored by 6 people. Awaiting for Putin to call in another distraction from TrumpRussia investigation.
Twitter Trend	Bot: Do you want me to talk about the top trending hashtags on twitter, or would you rather tell me a topic to focus on? User: I want to hear the top trending hashtags. Bot: Okay! This tag – scientologytheaftermath – is pretty popular. Want to hear a tweet from it? User: Yes. Bot: Great! Here's a tweet favored by 7090 people. Your support has made it possible for more people to come forward. We thank you. Want to hear another tweet from it? User: No
AIMLbot	Bot: How are you doing today? User: I'm not doing so well. Bot: Oh, I'm sorry to hear that. Hopefully I can make your day a bit brighter.
Chitchat	User: How are you? Bot: I am in NYC.
Horoscope	Bot: Tell me what your sign is. If you want to find it out, just say help. User: Help Bot: Tell me which month were you born in? User: October Bot: And which day in this month were you born? User: The twenty-fourth Bot: If your birthday is October the twenty-fourth, then you sign should be Scorpio. (horoscope) and would you like to know your match or rating? User: Both Bot: (matches and ratings)
Riddle	Bot: Let's solve some riddles, answers will be just one word. What has a neck but no head? User: I don't know Bot: Ok, the answer is bottle, there's a neck to a bottle, but no head. If two's company and three's a crowd, what are four and five? User: Nine. Bot: Correct! ... (more riddles)
Text Adventure Ghost Adventure Word Game Message Board	Descriptions for those modules have insufficient space to include, will be release on our website: <a href="http://dk00000000.github.io/wise_macaw/index.html">http://dk00000000.github.io/wise_macaw/index.html</a>