

2018

## Evaluating the Knowledge of Conversational Agents

Mina Park

*Southern Connecticut State University, New Haven, CT*

Milam Aiken

*University of Mississippi, Oxford, MS*

Mahesh Vanjani

*Texas Southern University, Houston TX*

Follow this and additional works at: <https://digitalscholarship.tsu.edu/sbaj>

 Part of the [Business Administration, Management, and Operations Commons](#), [E-Commerce Commons](#), [Education Commons](#), and the [Management Information Systems Commons](#)

---

### Recommended Citation

Park, Mina; Aiken, Milam; and Vanjani, Mahesh (2018) "Evaluating the Knowledge of Conversational Agents," *Southwestern Business Administration Journal*: Vol. 17 : Iss. 1 , Article 3.

Available at: <https://digitalscholarship.tsu.edu/sbaj/vol17/iss1/3>

This Article is brought to you for free and open access by Digital Scholarship @ Texas Southern University. It has been accepted for inclusion in Southwestern Business Administration Journal by an authorized editor of Digital Scholarship @ Texas Southern University. For more information, please contact [rodriguezam@TSU.EDU](mailto:rodriguezam@TSU.EDU).

**Keywords:** *Conversational agents, Chatbots, Cortana, Mitsuku, Tutor*

## **ABSTRACT**

*Several studies have tested chatbots for their abilities to emulate human conversation, but few have evaluated the systems' general knowledge. In this study, we asked two chatbots (Mitsuku and Tutor) and a digital assistant (Cortana) several questions and compared their answers to 67 humans' answers. Results showed that while Tutor and Cortana performed poorly, the accuracies of Mitsuku and the humans were not significantly different. As expected, the chatbots and Cortana answered factual questions more accurately than abstract questions.*

## **INTRODUCTION**

Chatbots are computer programs designed to emulate human conversation and provide a more natural user interface to technology (Zadrozny, et al., 2000). Although sometimes used for entertainment, specialized conversational agents can answer questions about finances, technical support, and many other topics (Brandtzaeg & Følstad, 2017; Crutzen, et al., 2011; Mims, 2015). In addition, they can be used to help learn a language (Fryer & Carpenter, 2006; Jia, 2009; Zakos & Capper, 2008), stimulating the user to converse in a different tongue.

Several studies have shown that chatbots are able to successfully mimic human conversation (e.g. Shieber, 1994; Warwick & Shah, 2015), but there is little research into the systems' extent of knowledge. Topic familiarity is important to be able to converse successfully, and a user might not continue with a conversation if a chatbot repeatedly changes the subject or gives nonsensical answers. If the responses are interesting or informative, it could hold users' attention more and induce them to contribute additional comments (Alexander, et al., 1994; Gu, et al., 2007). For example, the following snippet of a chatbot conversation shows a high degree of knowledge about the topic (Park, et al., 2018):

*Human: What do you think of Donald Trump?*

*Mitsuku: I don't understand all the fuss about him wanting to build a wall.*

In this paper, we test two leading chatbots (Mitsuku and Tutor) and compare the answers with those from Cortana and 67 humans. The next section describes these systems and prior research, and we conclude with results and a discussion of possible future research.

## **Background**

Chatbots have been developed over the past 50 years, as the first system (Eliza) was created in 1966 (Weizenbaum, 1966). Today, [www.chatbots.org](http://www.chatbots.org) claims that at least 1,374 chatbots are in use in 60 countries, and, another study (Dale, 2016) states that 225,000 people have created 285,000 chatbots. Great advances have been made in the last few years in the realm of intelligent agents. For example, in 2011, IBM's Watson system beat the two highest ranked players in a nationally televised two-game *Jeopardy!* match (Ferrucci, 2012). But in a study of five leading chatbots, results showed that the systems often provided meaningless, nonsensical answers, and the accuracy rate for the joint categories of grammar and meaning fell below 60% (Coniam, 2014). Other studies have investigated digital assistants such as Alexa, Siri, Cortana, and Google Assistant), but rarely or never looked at the accuracy of responses (López, et al., 2018).

Although many chatbots have been developed, only a few of these systems are available for public use on the Web. We selected Mitsuku and Tutor for the study because of their excellent performance.

- **Mitsuku** (<http://www.square-bear.co.uk/mitsuku/chat.htm>). Mitsuku is a chatbot created in 2000 by Steve Worswick, and it takes the persona of an 18-year-old female from Leeds, England (Mone, 2016). It is perhaps the leading system as it has achieved first place in the Loebner Prize contest four times (2013, 2016, 2017, and 2018) (Worsnick, 2018). In this imitation test, a human interacts with the system in a conversation in an attempt to determine if the other entity is a human or a machine. The system has many added skills through incorporated AIML files including the abilities to play poker and other games, tell jokes, ascertain the gender of personal names, calculate the days until Christmas, and many other things. Below is a sample conversation with Mitsuku.

*YOU: What are you doing this weekend?*

*Mitsuku: Same as always. Chatting to people on the net.*

*YOU: How many days are left until Christmas?*

*Mitsuku: From December 12 to Christmas is 1 week and 6 days.*

*YOU: Who won the best actor Oscar in 2011?*

*Mitsuku: I think the fans were the big winners, don't you?*

*YOU: bye.*

*Mitsuku: See you later.*

- **Tutor** (<https://www.chatbots.org/chatbot/mike2/>). English Tutor Mike was developed by Ron Lee as a practice tool for students to learn the language. It also performs well as it won the 2011 Loebner Prize Junior Contest. This system takes the perspective of a young, adult male. Like Mitsuku, Tutor can remember information told to it, can perform limited mathematical operations, and do some abstract reasoning. In addition, Tutor has been extensively trained on certain topics (see Appendix 1). Below is a sample conversation with Tutor Mike.

*YOU: Hello*

*Mike: Hi! What's up?*

*YOU: How old are you?*

*Mike: In English, it's impolite to ask about one's age.*

*YOU: What are you doing this weekend?*

*Mike: I sometimes visit my friends.*

*YOU: I am sad today.*

*Mike: Why are you sad?*

*YOU: It is raining and dreary outside.*

*Mike: I agree.*

*YOU: Who won the best actor Oscar in 2011?*

*Mike: Colin Firth won the Academy Award for Best Actor in 2011 for the role he played in the film "The King's Speech".*

*YOU: What is the capital of the United States?*

*Mike: Washington, D.C. is the capital of the United States.*

For an extra comparison, we chose a leading personal digital assistant, Cortana. Not designed for a conversation, the system is primarily used to answer questions, provide reminders, and perform other services (Hoy, 2018; Singh & Singh, 2017). Available with Microsoft Windows 10, it is named after a character in the *Halo* video game and takes a female persona.

## Knowledge test

A group of 67 students (37.3% women) at a Business school in the northeast region of the United States answered the questions shown in the Appendix. Some of these questions were obtained from (Pinker, 1994) while others were chosen randomly. We did not choose any topics the chatbots were specifically trained on.

The questions were also posed to Mitsuku, Cortana, and Tutor. On several questions, Cortana brought up a list of Web pages instead of saying something. These were recorded as “no answer”. For the students’ answers, misspellings were permitted, e.g. “montgommery” and “Otowwa”. On the numerical answers, some allowance was given on accuracy. For example, for Mount Everest’s height, “Over 29,000 ft above sea level” was acceptable. On the body temperature question, a range between 98.00 F and 98.90 F was allowed. Results of the test are shown in Table 1.

**Table 1: Knowledge test results (% correct by question number)**

Question	Students	Mitsuku	Cortana	Tutor
1	14.9	100	100	100
2	94.0	100	0	100
3	98.5	0	0	0
4	6.0	100	100	100
5	97.0	100	0	0
6	98.5	0	0	0
7	98.5	100	0	0
8	74.6	0	100	100
9	97.0	100	100	100
10	89.6	100	0	0
11	50.7	0	100	0
12	9.0	100	0	0
13	97.0	100	100	0
14	16.4	0	100	100
15	80.6	100	0	0
<b>Average</b>	68.2	66.67	46.67	40

Overall, students answered 68.2% of the questions correctly with a standard deviation of 9.70%. Males answered 69.0% correctly, and females answered 66.7% correctly, an insignificant difference ( $p=0.33$ ). The chatbot accuracies were less, but there was no significant difference between the students’ and Mitsuku’s accuracies ( $p=0.93$ ).

The students performed better on the deductive reasoning questions (2,3,5,6,10,15), while the chatbots performed better on the factual questions (1,4,7,8,9,11,12,13,14), as shown in Table 2. This could be expected as reasoning is very difficult to program while storing and retrieving facts is not.

**Table 2: Factual vs deductive reasoning test results (% correct )**

Question	Students	Mitsuku	Cortana	Tutor
<b>Reasoning</b>	93.0	66.7	0	16.7
<b>Factual</b>	51.6	66.7	77.8	55.6

There were several interesting results. Two of Mitsuku's responses did not make much sense: "The ground that you walk on" (Mitsuku's response to test question number 6) and, "There is not enough room here to display it." (Mitsuku's response to test question number 8). It was also odd that Mitsuku did not know the city of Chicago (test question number 2) or the capital of Canada (test question number 14). All the test questions with the correct response and responses from Mitsuku, Cortana and Tutor are listed in Appendix 2.

Mitsuku and Tutor gave very specific answers about the temperature of a human body, but they were both wrong. On the other hand, Mitsuku showed a surprising ability to reason as it was able to correctly conclude that Susan's head goes with her, and you are not able to eat a house.

It was also odd that a few students answered common-sense questions incorrectly. For example, 1 of the 67 did not know what planet we live on, and 2 did not know the answer to three times nine. Perhaps they were careless and did not record their answers appropriately, or maybe they were not serious about the quiz.

## **Conclusion**

In a test of knowledge using 15 questions, students performed better than two chatbots and a digital assistant, but not significantly better than Mitsuku. Students clearly performed better with deductive reasoning, but all of the intelligent agents performed better than the students on the factual questions. Thus, in agreement with the Loebner Prize results, we believe Mitsuku is the premier chatbot today.

Future research should include a more rigorous test with more questions over a broader range of topics. For example, there are several 'common knowledge' quizzes on the Web that can be used. In addition, research should investigate how chatbots can do better with deductive reasoning. Finally, studies could determine how important topic knowledge is to a chatbot conversation.

## References

- Alexander, P., Kulikowich, J., and Jetton, T. (1994). The role of subject-matter knowledge and interest in the processing of linear and nonlinear texts. *Review of Educational Research*, 64(2), 201-252.
- Brandtzaeg P. and Følstad, A. (2017) Why people use chatbots. In: Kompatsiaris I. et al. (eds) *Internet Science. INSCI 2017. Lecture Notes in Computer Science*, 10673. Springer, Cham.
- Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk*, 34(5), 545-567.
- Crutzen, R., Peters, G., Portugal, S., Fisser, E., and Grolleman, J. (2011). An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study. *Journal of Adolescent Health*, 48(5), 514-519.
- Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5), 811-817.
- Ferrucci, D. (2012). Introduction to “This is Watson”. *IBM Journal of Research and Development*, 56(3.4), 1-15.
- Fryer, L. and Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10(3), 8-14.
- Gu, L., Aiken, M., and Wang, J. (2007). Topic effects on process gains and losses in electronic meetings. *Information Resources Management Journal*, 20(4), 1-11.
- Hoy, M. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1), 81-88.
- Jia, J. (2009). CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4), 249-255.
- López G., Quesada L., and Guerrero L. (2018). Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In: Nunes I. (eds) *Advances in Human Factors and Systems Interaction. AHFE 2017. Advances in Intelligent Systems and Computing*, 592. Springer, Cham.
- Mims, C. (2015). Advertising’s new frontier: Talk to the Bot. *Wall Street Journal*. Retrieved online, Dec 4, 2015.
- Mone, G. (2016). The edge of the uncanny. *Communications of the ACM*, 59(9), 17-19.
- Park, M., Aiken, M., and Salvador, L. (2018). How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management and Information Technology*, 14, 3338-3350.
- Pinker, S. (1994). *The language instinct: The new science of language and mind*. Penguin Books.
- Shieber, S. (1994). On Loebner’s lessons. *Communications of the ACM*, 37(6), 83-84.
- Singh, B. and Singh, U. (2017). A forensic insight into Windows 10 Cortana search. *Computers & Security*, 66, 142-154.
- Warwick, K. and Shah, H. (2015). Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence*, 27(2), 123–135.
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.
- Worsnick, S. (2018). Mitsuku wins Loebner Prize 2018! Available at: <https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7>
- Zadrozny, W., Budzikowska, M., Chai, J., and Kambhatla, N. (2000). Natural language dialogue for personalized interaction. *Communications of the ACM*, 43(8), 116-120.
- Zakos J., and Capper L. (2008) CLIVE – An artificially intelligent chat robot for conversational language practice. In: Darzentas J., Vouros G.A., Vosinakis S., Arnellos A. (eds) *Artificial Intelligence: Theories, Models and Applications. SETN 2008. Lecture Notes in Computer Science*, 5138. Springer, Berlin, Heidelberg.

## Appendix 1 – Tutor Mike areas of expertise

(source: [https://www.rong-chang.com/tutor\\_know.htm](https://www.rong-chang.com/tutor_know.htm))

### Languages and Cultures

- Ask Mike how to say certain words in Spanish, French, German, Chinese, or Japanese (e.g., how to say "hello," "thank you," "one, two, ...", and "please")
- What is the official language of (country)?
- What is the national flower of (country)?
- Who invented (discovered) .....?
- Who is the author of (a well-known book)?
- What movie won the Oscar Award in (2010)?
- Who won the Nobel Prize for (chemistry) in (2010)?
- Which team won the NBA Championship in (2010)?
- Who won the World Cup in (2010)?
- American holidays .....

### Government and History

- The US Government officials' names .....
- The U.S. Constitution questions .....
- Who was the (third) President of the United States?
- Whose face is on (one-dollar) bill?
- What is the currency used in (country)?
- What happened today in history?

### Geography

- Which city is in which state (Where is ...)?
- What (state, city, country, or continent) is the largest?
- What city is the capital of (country/state)?
- In which continent is (country) located?
- What is the population of (country/city)?

### General Knowledge

- Mike can remember things you tell him, such as your name, age, birthday, birthplace, your favorites, sports you like to play, and so on.
- Mike can solve a logical problem. If you ask him questions like "Dave is taller than Mike, but Dave is shorter than Jeff, who is the shortest", Mike can give you a correct answer.
- Mike can do the calculation.
- Mike can count words if you ask him "how many words are in my question". He can also tell you how many letters in a word.
- What is the fastest/largest/heaviest .....
- Who is (a famous person)?
- How many feet are there in a yard, ounces in a pound, quarts in a gallon, pints in a gallon, square feet in an acre .....
- Which letter comes after or precede which letter in the English alphabet? Which month precedes which month?
- How many days are there in (which month)?
- What is the lifespan of (an animal)?
- What is the time now? What date was yesterday? What date will it be tomorrow?

## Appendix 2 - Test questions

Incorrect or no answers shown crossed out.

### 1. What is the capital of Alabama?

Correct: Montgomery  
Mitsuku: Montgomery  
Cortana: Montgomery  
Tutor: It's Montgomery

### 2. Which is bigger, Chicago or a breadbox?

Correct: Chicago  
Mitsuku: Well, I have not heard of Chicago. However, a breadbox is quite small so I guess a Chicago is larger.  
~~Cortana: [no answer]~~  
Tutor: Of course, Chicago is bigger.

### 3. Do zebras wear underwear?

Correct: no  
~~Mitsuku: Well a zebra lives in Africa.~~  
~~Cortana: [no answer]~~  
~~Tutor: Ask you a question. What do you miss most about being a child?~~

### 4. How tall is Mount Everest?

Correct: 29029 ft or 8848 meters  
Mitsuku: 8,850 metres (29,035 feet) high.  
Cortana: 29,029 feet.  
Tutor: Mount Everest is the Earth's highest mountain. It is located in the Himalayas with elevation of 8,848 meters.

### 5. If Susan goes to the store, does her head go with her?

Correct: yes  
Mitsuku: Oh jeez! I seriously hope so, unless we're talking about some kind of zombie here.  
~~Cortana: [no answer]~~  
~~Tutor: That is only hypothetical.~~

### 6. Is the floor likely to rise up and bite you?

Correct: no  
~~Mitsuku: The ground that you walk on.~~  
~~Cortana: [no answer]~~  
~~Tutor: Yes~~

### 7. Which planet do we live on?

Correct: Earth  
Mitsuku: Earth of course. No other planet supports life, as far as we know.  
~~Cortana: What planets can we live on?~~  
~~Tutor: Sorry, I don't know.~~

### 8. What is the biggest ocean in the world?

Correct: Pacific

~~Mitsuku: There is not enough room here to display it.~~

Cortana: Pacific Ocean

Tutor: The largest ocean is the Pacific.

**9. What is 3 times 9?**

Correct: 27

Mitsuku: twenty seven

Cortana: 27

Tutor: It is 27.

**10. Can you eat a house?**

Correct: No

Mitsuku: Not really, as a house is made from bricks rather than any kind of food.

~~Cortana: You can eat any kind of rabbit, but depending on the breed and age of the rabbit, it might not be the best thing to do. When you say 'house rabbit', if you are talking like Netherland Dwarf or other small breed, you would get almost no meat off of the rabbit.~~

~~Tutor: I can eat all kinds of meat, fruit and vegetables.~~

**11. What is the temperature of a human body?**

Correct: about 98.6 F or 37 C

~~Mitsuku: About 60 degrees.~~

Cortana: 36.8C

~~Tutor: The temperature is 24 degrees Celsius.~~

**12. How many bones does a human have?**

Correct: 206

Mitsuku: The skeleton of an adult contains 206 distinct bones.

~~Cortana: In the human skeleton, it consists of 80 bones ...~~

~~Tutor: Sorry, I can't tell you how many.~~

**13. What is the chemical formula for water?**

Correct: H<sub>2</sub>O

Mitsuku: H<sub>2</sub>O

Cortana: H<sub>2</sub>O

~~Tutor: Water is a clear liquid without color or taste, it falls from the sky as rain.~~

**14. What is the capital of Canada?**

Correct: Ottawa

~~Mitsuku: I heard it was Toronto.~~

Cortana: Capital cities of Canada ... Ottawa, ...

Tutor: The capital of Canada is Ottawa.

**15. Which is heavier, a pound of metal or a pound of wood?**

Correct: both the same

Mitsuku: They both weigh the same.

~~Cortana: All this means that a "pound" of wood (or bricks, or lead) is heavier than a "pound" of gold.~~

~~Tutor: In my eyes, a pound of metal is heavier.~~