PEEP-Talk: A Situational Dialogue-based Chatbot for English Education

Seungjun Lee[1], **Yoonna Jang**[1], **Chanjun Park**[1,2], **Jungseob Lee**[1]
Jaehyung Seo[1], **Hyeonseok Moon**[1], **Sugyeong Eo**[1], **Seounghoon Lee** [3]
Bernardo Nugroho Yahya[4], **Heuiseok Lim**[1*]
[1]Korea University, South Korea, [2]Upstage, South Korea
[3]Institute for Infocomm Research, A*STAR, Singapore
[4]Hankuk University of Foreign Studies, South Korea
{dzzy6505, morelychee, bcj1210, omanma1928, seojae777, glee889, djtnrud, limhseok}@korea.ac.kr
chanjun.park@upstage.ai, lees1@i2r.a-star.edu.sg, bernardo@hufs.ac.kr

Abstract

English is acknowledged worldwide as a mode of communication. However, due to the absence of realistic practicing scenarios, students learning English as a foreign language (EFL) typically have limited chances to converse and share feedback with others. In this paper, we propose PEEP-Talk, a real-world situational dialogue-based chatbot designed for English education. It also naturally switches to a new topic or situation in response to out-of-topic utterances, which are common among English beginners. Furthermore, PEEP-Talk provides feedback score on conversation and grammar error correction. We performed automatic and user evaluations to validate performance and education efficiency of our system. The results show that PEEP-Talk generates appropriate responses in various real-life situations while providing accurate feedback to learners. Moreover, we demonstrate a positive impact on English-speaking, grammar, and English learning anxiety, implying that PEEP-Talk can lower the barrier to learning natural conversation in effective ways.

1 Introduction

In the era of globalization, English is used as a worldwide international language (Kramsch, 2014). A number of countries have incorporated the acquisition of foreign language communication skills into their lifelong learning objectives (Luna Scott, 2015). Altalbe and Wilkinson (2013) have identified several areas, including education, tourism, and business, where direct communication with people is crucial, and conversation skill is considered as the most essential of the various language competencies.

However, students learning English as a foreign language (EFL) do not have sufficient opportunities to practice real-life English conversations (Jdetawy, 2011). To address this issue, recent technologies such as smart speakers and conversational models have been applied in English education (Tai and Chen, 2020; Alsadoon, 2021; Li et al., 2017). In particular, chatbots have shown promising results in improving the communication skills and learning motivation of students (Fryer and Carpenter, 2006).

Despite their potential, existing chatbot-based educational platforms face several challenges in providing effective language learning experiences. These chatbots employ hand-crafted and pattern-matching rules, limiting their communication ability and responding appropriately to out-of-situation utterances (Tyen et al., 2022; Kim et al., 2019). Furthermore, as smart speakers and conversational AI models are not fully considered for educational purposes (Terzopoulos and Satratzemi, 2020; Ji et al., 2022), they cannot cover various topics in real-life activities or provide educational feedback.

To address these challenges, we propose PEEP-Talk, a situational dialogue-based chatbot for English education. It consists of a conversation module, context detector (CD), and grammar error correction (GEC) modules. The conversation module generates proper utterances considering the given situations with our proposed dataset, called SITUATION-CHAT. It covers a variety of real-world situations. To address the previous chatbots' inability to interact with dialogue out of topic or situation, the CD module changes the situation when the conversation digresses from the current situation. PEEP-Talk also provides feedback on learners' utterances with the situation similarity score, linguistic acceptability score, and grammar error correction.

We quantitatively verified the performance of each module of PEEP-Talk and conducted a user study to verify its effectiveness in English education in a real-world environment. To the best of our knowledge, there have been few attempts in NLP research to conduct a user study that verifies the performance and satisfaction of integrated modules. The comprehensive evaluation of PEEP-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 190–207
July 10-12, 2023 ©2023 Association for Computational Linguistics

Talk demonstrates the potential of our situational dialogue-based chatbot and feedback approach for effective English education. We deploy our methodology on a demo site. The code and dataset[1] are publicly available to facilitate further studies.

2 Chatbots for English Education

The field of language learning has been revolutionized by the emergence of chatbots. While previous computer-assisted language learning (CALL) systems have provided lessons and practice in writing, vocabulary, grammar, and other areas, they have limitations in that learners eventually need an actual human, such as a teacher or a colleague, to practice conversation (Chapelle, 2005). Chatbots have the potential to bridge this gap in CALL systems, as they offer more natural conversation and feedback similar to that of human instructors (Kim et al., 2019).

Commercial chatbots for English education have become increasingly popular, including Andy, Mondly Speak Now, Duolingo, and Babbel. Andy is a virtual tutor application that aids users in learning English, while Mondly features language lessons in various situations with a virtual teacher. Speak Now uses AI-driven virtual avatars as instructors for learning English. Duolingo offers gamified language lessons and has incorporated chatbots into its system, while Babbel features chatbots as part of its teaching methodology. The features of these existing chatbots for English education are also shown in Table 1.

However, one of the main problems with current English education chatbots is that they are unable to provide a genuinely free and natural conversation experience. Many chatbots rely on rule-based systems that are limited in generating diverse and spontaneous responses to user input (Fryer and Carpenter, 2006; Kim et al., 2019). As a result, learners often feel frustrated and disconnected from the language they are trying to learn (Fryer and Nakao, 2009). Additionally, these chatbots may struggle to understand idiomatic expressions, provide personalized feedback, and recognize situational contexts. Without engaging in meaningful and contextually relevant conversations, learners may struggle to develop the communication skills necessary for real-life interactions in English-speaking environments.

	features			URL
	AI-based	Situation	Feedback	
Andy	✗	✗	✗	`andychatbot.com`
Mondly	✗	✗	✗	`mondly.com`
Speak Now	✗	✓	✗	`speaknow.ai`
Duolingo	✓	✗	✓	`duolingo.com`
Babbel	✗	✗	✓	`babbel.com`
PEEP-Talk	✓	✓	✓	`peeptalk.us`

Table 1: Comparison of chatbots for learning English. AI-based (not rule-based), Situation (use of situational dialogues), and Feedback (provision of learner feedback)

3 PEEP-Talk

We introduce PEEP-Talk and discuss its motivation and unique features in §3.1. Then, we present the SITUATION-CHAT in §3.2, which contains diverse and situational expressions. The overall architecture of PEEP-Talk, including its three modules - conversation, context detector (CD), and GEC, is described in §3.3. Finally, §3.4 covers PEEP-Talk's deployment.

3.1 Why PEEP-Talk?

While existing chatbot-based educational platforms have shown promising results in improving students' communication skills and motivation to learn English (Jia and Ruan, 2008; Fryer and Carpenter, 2006; Haristiani, 2019), they still face several challenges in providing a practical language learning situation. One of the significant limitations of these chatbots is their inability to interact with utterances out of topic or situation. Additionally, smart speakers and conversational AI models are not fully optimized for educational purposes, as they cannot cover various topics in real-life activities or provide feedback on grammar errors.

To address these challenges, we propose PEEP-Talk, a situational dialogue-based chatbot for English education. PEEP-Talk is designed to generate contextually appropriate utterances using our proposed dataset, called SITUATION-CHAT. Furthermore, the context detector (CD) module enables PEEP-Talk to identify and adapt to changes in conversation topics, providing a more natural and engaging learning experience. The grammar error correction (GEC) module provides instant feedback to learners to improve their linguistic accuracy.

PEEP-Talk's situational dialogue-based approach and dataset with diverse situations offer an effective language learning experience for EFL

---

[1] `https://github.com/metterian/peep-talk`

learners. PEEP-Talk aims to address the limitations of existing chatbot-based educational platforms and provide an interactive and engaging English education environment.

3.2 The SITUATION-CHAT Dataset

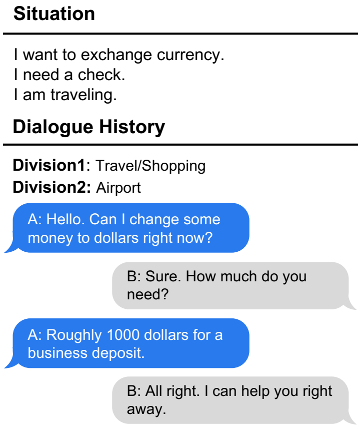
This section presents the SITUATION-CHAT dataset, a situational dialogue-based dataset for English learning. An example of this dataset is shown in Figure 1. The dataset contains various situation dialogue such as asking for directions, talking with friends in school, and company interviews, among others. To construct our dataset, we follow the definition of a situational dialogue in Klinghoffer (2008). Situational dialogue is a learning approach where learners participate in role-playing through a routine activity, allowing for a more natural and engaging conversation. This approach has been shown to enhance the learners' communication skills, as it allows them to practice real-life scenarios (Klinghoffer, 2008).

To construct our proposed dataset, We adopted AI hub [2]'s dialog corpus, a Korean-English translation corpus that contains conversations in a variety of contexts across multiple domains. This corpus is composed of division (domain), predefined situations, and dialogue history. To develop a situational dialogue-based conversational AI, the contextual information of predefined situations is insufficient. Therefore, we employ human annotators to write additional descriptions for the predefined situations. A details of this process is provided in Appendix A.

The dataset contains 16,298 dialogues with 65,192 utterances for the training set and 1,000 dialogues with 4,000 utterances for the validation and test sets, respectively. Each turn in the dataset has an average length of 14 words, and it includes 303 different situations covering various domains, including shopping, traffic, and travel. The statistical details of the dataset are written in Appendix A.3.

3.3 Overall Architecture

The architecture of PEEP-Talk consists of three modules: the conversation module, the context detector (CD) module, and the grammar error correction (GEC) module. The conversation module generates responses that are conditioned on the situation and the dialogue history. The CD module checks the user's input to determine its appropriateness in the topic or situation and provides feedback scores. The GEC module corrects grammatical errors in the conversation. Figure 2 provides an overview of the PEEP-Talk architecture.



Situation

I want to exchange currency.
I need a check.
I am traveling.

Dialogue History

Division1: Travel/Shopping
Division2: Airport

A: Hello. Can I change some money to dollars right now?

B: Sure. How much do you need?

A: Roughly 1000 dollars for a business deposit.

B: All right. I can help you right away.

Figure 1: Example of the SITUATION-CHAT.

3.3.1 Conversation Module

Compared to existing rule-based educational systems, the conversation module provides a more natural and diverse conversation experience to learners by employing DialoGPT (Zhang et al., 2019), a pre-trained language model trained on dialog corpus. It is further fine-tuned on our SITUATION-CHAT for learning situational expressions following the methodology presented in Wolf et al. (2019b). The input to the conversation module includes the situation, dialogue history, and candidate answers.

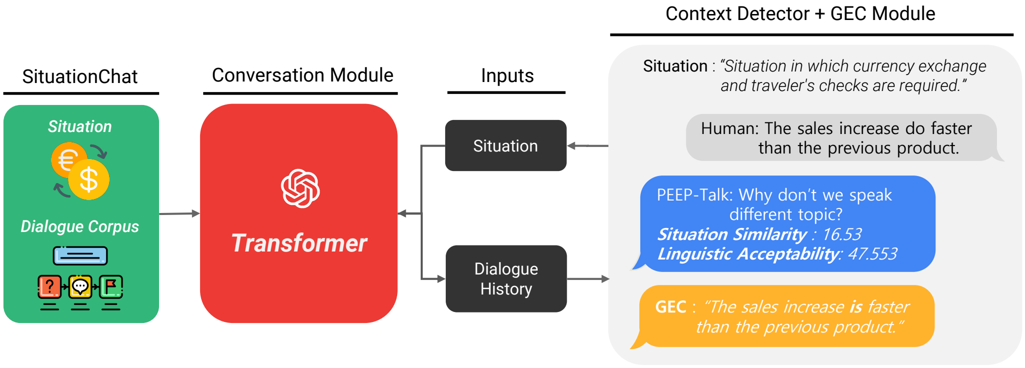
3.3.2 Context Detector Module

The primary objective of CD module is to provide feedback on learners' responses when the conversation has deviated from the given situation. The main challenge that commercial chatbots face for English education is their inability to respond to inappropriate input from English beginners, including EFL learners (see Appendix H). Beginners often make linguistic errors or speak topics or situations that are out of topic in their conversations (Al Hosni, 2014; Sermsook et al., 2017), which the commercial chatbots are not equipped to handle appropriately. The CD module addresses this problem by checking the user's utterance to determine its appropriateness in the topic or situation and whether it is linguistically natural.

The CD module gives situation similarity and linguistic acceptability scores on dialogue and switches dialogue situations. The situation similarity score measures the similarity of the user's input

---

[2] The AI Hub, which can be accessed at `https://aihub.or.kr` is a public data platform operated by the government.



Situation : *"Situation in which currency exchange and traveler's checks are required."*

Human: The sales increase do faster than the previous product.

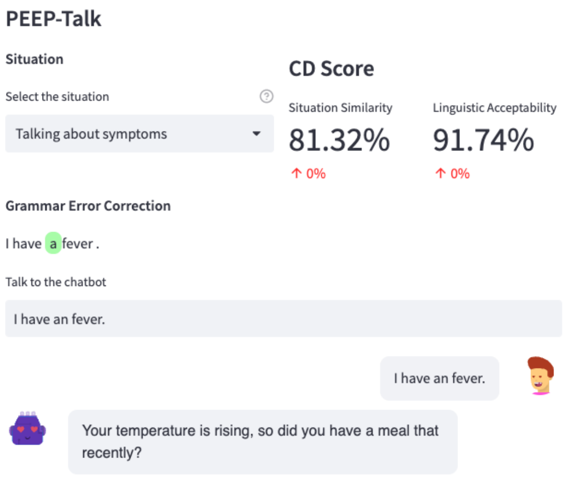
PEEP-Talk: Why don't we speak different topic?
Situation Similarity *: 16.53*
Linguistic Acceptability*: 47.553*

GEC : *"The sales increase* ***is*** *faster than the previous product."*

Figure 2: An overview of the modules in PEEP-Talk.

Models	**Validation PPL**	**Validation Hits@1**	**Test PPL**	**Test Hits@1**
Profile Memory	42.79	0.078	45.02	0.069
Lost In Conversation	-	0.168	-	0.152
GPT$_{SMALL}$	12.41	0.849	12.74	0.839
GPT-2$_{SMALL}$	12.50	0.839	12.56	0.848
DialoGPT$_{SMALL}$	12.35	0.850	12.55	0.856
DialoGPT$_{MEDIUM}$	14.77	0.884	13.89	0.864
DialoGPT$_{LARGE}$	**11.15**	**0.889**	**12.04**	**0.877**

Table 2: Experimental results for validation and test set of SITUATION-CHAT. PPL denotes perplexity and Hits@1 denotes the correct identification of the gold answer among 19 randomly sampled utterance candidates.


PEEP-Talk

Situation

Select the situation

Talking about symptoms

CD Score

Situation Similarity
81.32%
↑ 0%

Linguistic Acceptability
91.74%
↑ 0%

Grammar Error Correction

I have a fever .

Talk to the chatbot

I have an fever.

I have an fever.

Your temperature is rising, so did you have a meal that recently?

Figure 3: A screenshot of our demo website. Learners can choose various situations from the select box and receive responses. Grammatical errors are corrected and highlighted. The feedback scores are shown in the CD Score.

to the current situation (detailed in Appendix B), while the linguistic acceptability score checks the grammatical correctness of the user's input. The CD module utilizes XLNet (Yang et al., 2019), fine-tuned on MRPC (Dolan and Brockett, 2005) and CoLA datasets (Wang et al., 2019), respectively.

The CD module suggests a new situation if the situation similarity score is under a certain threshold. Furthermore, as shown in Figure 2, learners can see the feedback score of the CD module during the conversation. Switching situation and real-time feedback helps to enhance the learner's understanding of appropriate language use in different situations and promote their overall communication skills.

3.3.3 GEC Module

To provide accurate and efficient English grammar feedback to English learners, a deep learning-based GEC module has been integrated into PEEP-Talk in the form of a REST API. Specifically, we have adopted the approach described in Park et al. (2020), which uses a sequence-to-sequence (S2S) model. This approach is based on a noise sequence to denoise sequence (NS2DS) model, incorporating the copy mechanism (Gu et al., 2016). Notably, this method has been validated using commercial data and has been successfully applied in actual services. With this method, our GEC module ensures accurate and helpful real-time feedback on the learners' grammar errors.

3.4 Deployment

The proposed PEEP-Talk is deployed through the web to render it easier for more people to access our system. We used HuggingFace (Wolf et al., 2019a) and TextAttack (Morris et al., 2020) library for developing modules. We use FastAPI to make all our modules as a REST API, then we build the web application[3] with the Streamlit library. A screenshot of our demo website is shown in Figure 3.

---

[3] `http://peeptalk.us`

Model	Fluency	Situation Consistency	Engagingness
Human	4.54	4.51	4.63
Profile Memory	2.78	1.79	2.25
GPT-2$_{SMALL}$	2.54	3.11	3.44
DialoGPT$_{LARGE}$	**3.78**	**3.92**	**4.13**

Table 3: Results of human evaluation of conversation module. Human indicates the gold utterance of the test set annotated by humans.

4 Experiments

In this section, the effectiveness of the conversation module is verified through several automatic evaluation metrics and human evaluation. Also, we analyze CD and GEC modules in Appendix D and Appendix D.3 respectively.

4.1 Verification of Conversation Module

4.1.1 Quantitative Analysis

Method To empirically verify the conversation module in situational dialogue, we compare existing baselines and our model with the ability to generate and select the proper utterances. We utilize Profile Memory (Dinan et al., 2019) and Lost In Conversation as our baselines. For our model, GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2019) are exploited and fine-tuned on our data. Experimental settings are described in Appendix C. Furthermore, perplexity, and Hits@1 score indicating the correct identification of gold answers among 19 randomly sampled utterance candidates are used.

Results The experimental results of the conversation module are presented in Table 2. The experiment is conducted on validation and test sets of SITUATION-CHAT. DialoGPT$_{LARGE}$ exhibits the best performance with a perplexity of 12.04 and Hits@1 score of 0.877. This suggests that DialoGPT is more effective in generating and selecting appropriate responses in situational dialogue than the other models and baselines. Therefore, the results suggest that DialoGPT can be a promising model for situational dialogue. The experimental settings and details are provided in Appendix C.

4.1.2 Qualitative Analysis

Method We evaluate the effectiveness of PEEP-Talk in providing contextualized responses in various situations properly. The data of 30 situations are randomly selected from the test set of SITUATION-CHAT. For comparison with existing methodologies, we select DialoGPT$_{LARGE}$ fine-tuned on our proposed dataset, DailyDialog (Li et al., 2017), and the English education chatbots Andy and Speak Now.

Results Table 4 presents the results of the qualitative analysis. In the "Asking for directions" situation, PEEP-Talk gives a specific and helpful response, while DailyDialog's response lacks useful information. In "Requesting packaging", PEEP-Talk informs the customer about the availability of containers and demonstrates knowledge about the establishment, while DailyDialog's response is unhelpful. In the "Stopping credit card" situation, PEEP-Talk is straightforward in asking for necessary information, whereas Andy provides an unsolicited introduction. In "Pharmacy", PEEP-Talk provides a practical solution by recommending a medicine, while DailyDialog and Speak Now provide generic responses. PEEP-Talk shows that it can provide informative and helpful responses tailored to the situation at hand.

4.1.3 Human Evaluation

To validate the quality of the conversation module fine-tuned with SITUATION-CHAT, we conducted a human evaluation. First, fluency, situational consistency, and engagingness are measured as shown in Appendix E. We set the response level on a Likert scale (Likert, 1932) from 1 to 5. The evaluators are asked to evaluate the quality of the 150 randomly chosen model-generated utterances chosen from the test set of SITUATION-CHAT. The evaluators comprise English instructors, a teaching assistant, and three people with bachelor's degrees in English literature[4]. Table 3 reports the evaluation results. The result shows that DialoGPT$_{LARGE}$ exhibits the highest score in all factors and is the most similar to the human level, indicating the potential of our model for providing a more natural and engaging conversation experience for English learners.

5 User Evaluation

To verify the performance and satisfaction of PEEP-Talk as an English education tool, we conducted a user evaluation, which is rarely performed in NLP research. While other research often concludes their evaluation with human evaluation, we go further by actually operating our system and verifying its effectiveness in real-world English education.

---

[4] They were paid $0.4 per dialogue evaluation.

Situation & Previous utterance	**PEEP-Talk**	**DailyDialog**	**Andy**	**Speak Now**
Asking for directions "Can you tell me how to get to the nearest train station?"	"You can take the train to the City Hall Station if you go out exit number 2 on the 2nd."	"I can't, I'm not a train driver."	"Okay"	"Good Man."
Requesting packaging "Can I get the leftovers to go?"	"You can, but we don't have a container for soup dishes."	"I'm sorry, I don't have any leftovers."	"Sure, will you write me later?"	"I'm looking at you"
Stopping credit card "Hello? I lost my credit card and I am calling to report the loss."	"Yes, please tell me your social security number and the card number of the person who used it."	"I lost my credit card and I am calling to report the loss."	"I see"	"Let's keep going."
Pharmacy "I caught a bad cold. I cough a lot and have a high fever."	"Take this cold medicine pill three times a day."	"I hope you feel better soon!"	"Got it"	"Oh, You hurt my feelings"

Table 4: Comparison of PEEP-Talk with other conversational agents in different situations

		Group	**Mean**	**SD**	**Adj. Mean**	$F(1, 32)$	p
English Speaking	Accuracy	Exp.	2.41	0.51	2.58	0.77	0.18
		Ctrl.	2.67	0.62	2.64		
	Communication	Exp.	2.76	0.66	2.79	7.04*	0.03
		Ctrl.	2.93	0.70	2.69		
	Fluency	Exp.	2.93	0.62	2.81	3.6*	0.03
		Ctrl.	2.59	0.62	2.67		
	Comprehension	Exp.	3.35	0.49	3.47	3.98*	0.04
		Ctrl.	3.33	0.49	3.30		
	Maturity of the language	Exp.	2.71	0.59	2.98	12.52*	0.00
		Ctrl.	3.07	0.70	2.94		
	Grammar	Exp	72.22	8.95	64.51	2.78*	0.04
		Ctrl.	65.28	12.06	63.80		
	Learning anxiety	Exp	2.89	0.33	3.04	2.92*	0.04
		Ctrl.	3.08	0.61	3.14		

Table 5: Analysis of covariance of English-speaking performance, grammar, and learning anxiety for the two groups. Adj. Mean = adjusted mean. Exp. = experimental group with 18 people; Ctrl. = control group with 17 people. ${}^{*}p < 0.05$.

Method The users evaluation consisting of assessments in English speaking, grammar, English learning anxiety, and user satisfaction. In order to conduct the evaluation, we divided participants into experimental and control groups, with 18 and 17 people respectively. They perform pre-test and post-test before and after practicing with our system. The experimental group used PEEP-Talk for 30 minutes per day for a period of two weeks, while the control group did not use any language learning tools during the evaluation period. Participants in both groups were consisted of women in their twenties who are native Korean speakers and university graduates.They are composed of those who speak EFL. The details of the experimental design can be found in Appendix F.

Results Based on the analysis of covariance (ANCOVA) results in Table 5, the experimental group shows better performance in most dimensions of English speaking, except for accuracy, indicating that PEEP-Talk improves learners' overall speaking skills. In terms of grammar, the experimental group had a higher adjusted mean (64.51) compared to the control group (63.80), and the F-test result ($F(1, 32) = 2.78, p < 0.05$) indicates a significant difference in the grammar skills between the two groups. The English learning anxiety score was lower for the experimental group (adjusted mean = 3.04) than the control group (adjusted mean = 3.14). We conclude that PEEP-Talk has a positive impact on English-speaking, Grammar, and English learning anxiety after only two weeks of use.

5.1 PEEP-Talk Satisfaction Survey

We conduct satisfaction surveys to collect learners' satisfaction in a real English teaching environment. (presented in Appendix G). The survey results indicate that most users are satisfied with PEEP-Talk, with more than 70% of respondents finding the response of the chatbot natural and contextual. Moreover, 70 to 80% of respondents indicated that situational dialogue is educationally beneficial for conversational English and that grammar correction is helpful. However, the CD module receives a relatively low satisfaction rate of 46%, possibly due to the module switching the situation even when the user wished to continue talking about a specific situation.

6 Conclusion

In this paper, we propose PEEP-talk, an English education chatbot that allows learners to practice real-life conversations and receive feedback. Our research, pioneering in the realm of situational dialogue-based chatbots, was substantiated through rigorous qualitative and quantitative experiments. Furthermore, through user evaluations and satisfaction surveys, we confirmed that our method enhances educational efficiency and reduces learning anxiety for EFL learners in real-world English teaching environments. This research is pivotal as it introduces a novel approach to chatbot-based English education, enhancing learning efficiency and engagement. In the future, we will also integrate additional features, such as learner progress tracking and an administrative dashboard for personalized instruction.

Limitations

This study has certain limitations. Firstly, the effectiveness of the CD module in detecting inappropriate input needs improvement. Secondly, the current version of PEEP-Talk covers only a limited number of topics and situations, warranting an extension of the dialogue dataset. Thirdly, the observed improvement in English speaking ability may not be entirely attributable to PEEP-Talk, as external factors such as daily English speaking practice outside the experiment could also have contributed. Moreover, the participant pool was relatively small, suggesting a need for future studies with larger and more diverse groups. Lastly, PEEP-Talk's current limitation to text-based interactions highlights an area for development, as incorporating voice interactions could further enhance the user experience. Future research could also investigate the correlation between linguistic acceptability and speaking and grammar performance.

Ethical Statement

We uphold the ethical principles outlined by the ACL Code of Ethics in our research. The public dataset used in our study was obtained from a nationally operated and managed AI Hub[5]. Prior to training the DialogGPT model, we followed the preprocessing steps outlined in Zhang et al. (2019) to remove offensive language using syntax matching against large blacklists and excluded subreddits that are likely to contain offensive content. To further prevent generating offensive content, we will exclude tokens in a list of several stopwords using Hatebase[6]. Furthermore, we ensure the privacy of our users by not collecting any personal data and only providing responses based on the situations that the learners have selected. We prioritize the ethical considerations of our study to maintain the integrity of our research and prevent any potential harm.

Acknowledgments

We would like to express our deep appreciation for the generous support we have received for our research. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under grant number 2020-0-00368, titled "A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques". Additionally, this research was bolstered by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency under grant number R2020040045. Lastly, we gratefully acknowledge the support of the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01405), which was supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Samira Al Hosni. 2014. Speaking difficulties encountered by young efl learners. *International Journal on Studies in English Language and Literature (IJSELL)*, 2(6):22–30.

Reem Alsadoon. 2021. Chatting with ai bot: Vocabulary learning assistant for saudi efl learners. *English Language Teaching*, 14(6):135–157.

Ali A Altalbe and Brett Wilkinson. 2013. Designing games for learning english as a second language. In *2013 International Conference on Computer Applications Technology (ICCAT)*, pages 1–7. IEEE.

Carol A Chapelle. 2005. Computer-assisted language learning. In *Handbook of research in second language teaching and learning*, pages 767–780. Routledge.

Shu-Yun Chien, Gwo-Jen Hwang, and Morris Siu-Yung Jong. 2020. Effects of peer assessment within the context of spherical video-based virtual reality on efl students' english-speaking performance and learning perceptions. *Computers & Education*, 146:103751.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

---

[5] `https://aihub.or.kr`

[6] `http://hatebase.org`

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

L Fryer and K Nakao. 2009. Assessing chatbots for efl learner use. In *JALT2008 conference proceedings*, pages 849–857.

Luke Fryer and Rollo Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology*, 10(3):8–14.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Nuria Haristiani. 2019. Artificial intelligence (ai) chatbot as language learning medium: An inquiry. In *Journal of Physics: Conference Series*, volume 1387, page 012020. IOP Publishing.

Elaine K Horwitz, Michael B Horwitz, and Joann Cope. 1986. Foreign language classroom anxiety. *The Modern language journal*, 70(2):125–132.

Loae Fakhri Ahmad Jdetawy. 2011. Problems encountered by arab efl learners. *Language in India*, 11(3).

Hyangeun Ji, Insook Han, and Yujung Ko. 2022. A systematic review of conversational ai in language education: focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, pages 1–16.

Jiyou Jia and Meixian Ruan. 2008. Use chatbot csiec to facilitate the individual learning in english instruction: A case study. In *International conference on intelligent tutoring systems*, pages 706–708. Springer.

Na-Young Kim, Yoonjung Cha, and Hea-Suk Kim. 2019. Future english learning: Chatbots and artificial intelligence. *Multimedia-Assisted Language Learning*, 22(3):32–53.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Curtis L Klinghoffer. 2008. Situational dialogues in a community college: English as a second language curriculum. *Online Submission*.

Claire Kramsch. 2014. Teaching foreign languages in an era of globalization: Introduction. *The modern language journal*, 98(1):296–311.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Cynthia Luna Scott. 2015. The futures of learning 2: What kind of learning for the 21st century?

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Chanjun Park, Yeongwook Yang, Chanhee Lee, and Heuiseok Lim. 2020. Comparison of the evaluation metrics for neural grammatical error correction with overcorrection. *IEEE Access*, 8:106264–106272.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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.

Kanyakorn Sermsook, Jiraporn Liamnimitr, and Rattaneekorn Pochakorn. 2017. The impact of teacher corrective feedback on efl student writers' grammatical improvement. *English Language Teaching*, 10(10):43–49.

Tzu-Yu Tai and Howard Hao-Jan Chen. 2020. The impact of google assistant on adolescent efl learners' willingness to communicate. *Interactive Learning Environments*, pages 1–18.

George Terzopoulos and Maya Satratzemi. 2020. Voice assistants and smart speakers in everyday life and in education. *Informatics in Education*, 19(3):473–490.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A SITUATION-CHAT Dataset

A.1 Dialogue Corpus Collection

The dialogue corpus comprises division 1, division 2, predefined situations, and dialogue history. Divisions are domain information in a hierarchical structure, and predefined situation refers to a short description of the situation (e.g., "hospital admission situation", "car accident situation"). Dialogue history is composed 4-turns dialogue of speaker A and B in each situation. We only use English data from the Korean-English parallel dataset. Further information regarding this corpus is provided in Table 6.

With the aid of a professional English instructor, we select situations and dialogue histories from the dialogue corpus, such as "taking a taxi" or "asking to take a picture". Consequently, 330 situations out of 2,779 situations from dialogue corpus are selected with a total of 73,192 dialogue histories.

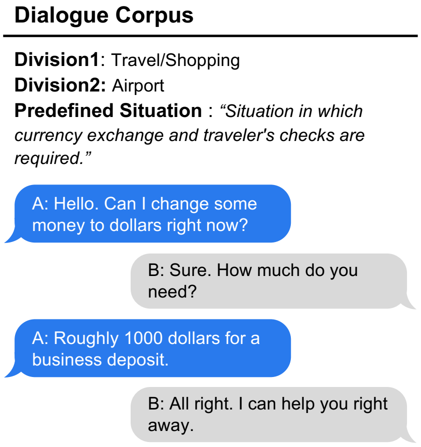
Division1	Domain information about dialogue sentences e.g.) Lectures, news, discussions, etc.
Division2	Specialty information to subdivide Division 1 e.g) Office, school, meetings, etc.
Predefined situation	One-line description of the situation e.g.) Exchange opinions
Dialogue history	Korean-English Conversation on Situations e.g.) How is the market's reaction to the released product.

Table 6: Information and examples of dialogue corpus.

A.2 Situation Annotation

As shown in Figure 4, predefined situation in the dialogue corpus includes only one sentence to explain the situation. Therefore, the situational information is augmented with various virtual descriptions. For example, to describe the situation of currency exchange as shown in Figure 5, the dialogue history and the information of the domains are used to generate the sentences that describe the given situation. With four human annotators[7], at least 2 and 5 sentences were generated for one predefined situation.

Specifically, we provide domain information, predefined situation, and conversation history and human annotators follow these rules when annotating the situations: (i) The subject of the sentence should be preferably in the first person. (ii) A description of the situation that should be inferred from the given information. In addition, they generated sentences in under 15 words.



Dialogue Corpus

Division1: Travel/Shopping
Division2: Airport
Predefined Situation : *"Situation in which currency exchange and traveler's checks are required."*

A: Hello. Can I change some money to dollars right now?

B: Sure. How much do you need?

A: Roughly 1000 dollars for a business deposit.

B: All right. I can help you right away.

Figure 4: Example of dialogue corpus.

Annotated situation

I want to exchange currency.
I need a check.
I am traveling.

Figure 5: Example of annotated situation.

A.3 Dataset Statistics

The statistics of the entire dataset and doamins are presented in Table 7 and Figure 6, respectively.

Category	Train	Valid	Test
Division 1	5	5	5
Division 2	38	32	30
# Situations	248	42	40
# Dialogues	16,298	1,000	1,000
# Utterances	65,192	4,000	4,000

Table 7: Statistics of SITUATION-CHAT. In case of the situations, they are separated into unseen ones for each other.

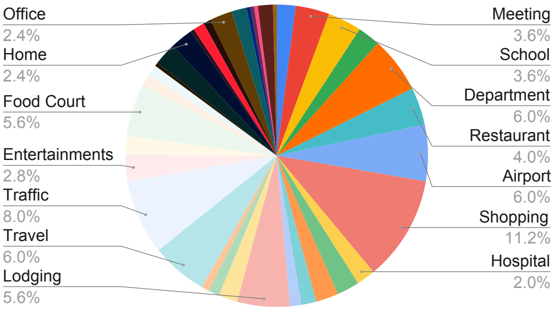


Figure 6: The domains of situations in SITUATION-CHAT.

---

[7] They were paid $0.2 per one situation; they had a Bachelor's degrees in English linguistics

B Situation Similarity Score

We utilize semantic textual similarity (STS) to measure the situation similarity score by comparing the user's input and utterances of the predefined situation to determine it is out of situation (topic). Each predefined situation in SITUATION-CHAT contains an average of 480 utterances. The STS scores are computed with the user's input and every utterance belong to the predfined situation dialogue data as shown in Figure 7. We use the maximum value of STS scores as the situation similarity score.

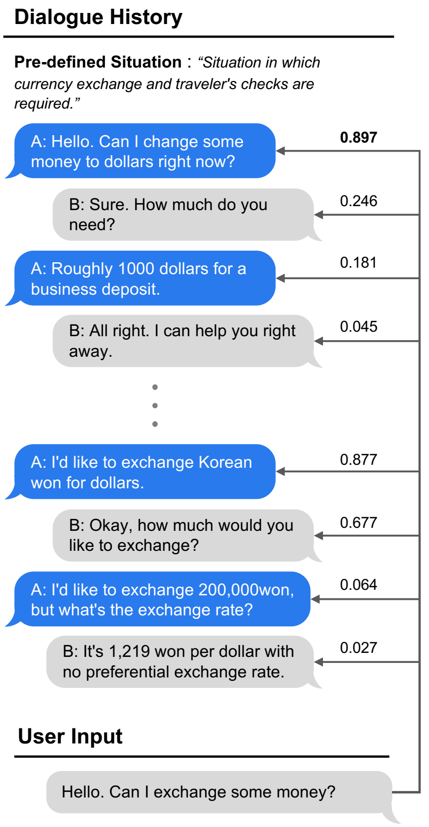

Dialogue History
Pre-defined Situation : "Situation in which currency exchange and traveler's checks are required."
A: Hello. Can I change some money to dollars right now?
0.897
B: Sure. How much do you need?
0.246
A: Roughly 1000 dollars for a business deposit.
0.181
B: All right. I can help you right away.
0.045
A: I'd like to exchange Korean won for dollars.
0.877
B: Okay, how much would you like to exchange?
0.677
A: I'd like to exchange 200,000won, but what's the exchange rate?
0.064
B: It's 1,219 won per dollar with no preferential exchange rate.
0.027
User Input
Hello. Can I exchange some money?

Figure 7: Example of scoring the situation similarity. We score the STS for pairs of the user's input and utterances of the predefined situation in SITUATION-CHAT, then we consider the maximum score as the situation similarity. In the figure, the score represents the STS.

C Experimental Setting of Conversation Module

The experimental models used are based on ConvAI2's methodologies, which are similar to the proposed approaches and publicly available. The baseline model of ConvAI2 is Profile Memory, while the state-of-the-art models are *TransferTransfo* and Lost In Conversation, which respectively achieved the best results in the automatic and human evaluation.

Fine-tuning Details The detailed fine-tuning setting in this study is as follows. We trained our model on one GeForce RTX 8000 GPU for 18 hours with batch size of 8 and 2 epochs following the fine-tuning method of *TransferTransfo*(Wolf et al. (2019b)). We set the random seed as 42. Further, the Adam optimizer (Kingma and Ba, 2014) is employed and with learning rate set to converge from $6.25e-5, \beta_1 = 0.9, \beta_2 = 0.999$ to 0. Further, all layers are set with a dropout of 0.1 and a ReLU activation function. In terms of the number of model parameters, GPT has 117M, GPT-2$_{SMALL}$ has 124M, DialoGPT$_{SMALL}$ has 124M, DialoGPT$_{MEDIUM}$ has 355M and DialoGPT$_{LARGE}$ has 774M of trainable parameters.

Decoding details Beam search with the size of 4 and N-gram filtering methods are used during the generation process. The final beam is ranked according to the scala combination of the length-normalized utterance probability and next-utterance classification score. In addition, increasing the importance of the next-utterance classification score produces utterances more closely related to the conditioned situation sentences; however, the diversity of the dialogue is reduced.

D Verification of CD Module

The performance of the feedback on dialogue and the switching situation of CD module is verified. The experimental results of the two functions are shown in Table 9. The linguistic acceptability of feedback on dialogue is estimated using the validation and test set of CoLA. Further, the performance of switching situations using situation similarity of CD module is evaluated based on CEEU. Specifically, pre-trained transformers are adopted, and a test set of SITUATION-CHAT is evaluated with 1,000 situations in the experiments.

D.1 Experimental Setting

The CD module is trained on two datasets: Microsoft research paraphrase corpus (MRPC) and corpus of linguistic acceptability (CoLA). The training set of MRPC consists of 3,260 pairs, while the validation and test sets consist of 408 pairs respectively. For CoLA, the training set consists of 7,551 sentences, while the validation and test sets consist of 1,000 sentences each. These datasets are used to evaluate the performance of the CD module in terms of situation similarity and linguistic acceptability.

We use TextAttack (Morris et al., 2020) for this experiment. We fine-tuned the MRPC dataset for situation similarity and the CoLA dataset for linguistic acceptability. We use pre-trained language models, which are BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). We set the seed as 42. The experimental settings are presented in Table 8.

	Model	Epoch	Batch	LR	SeqLen	# Params
Situation	BERT	5	16	2e-05	256	110M
	ALBERT	5	32	2e-05	128	12M
	RoBERTa	5	16	3e-05	256	125M
	XLNet	5	32	5e-05	256	117M
Linguistic	BERT	5	32	2e-05	128	110M
	ALBERT	5	32	3e-05	128	12M
	RoBERTa	5	32	2e-05	128	125M
	XLNet	5	32	5e-05	128	117M

Table 8: We denote situation similarity as Situation, linguistic acceptability as Linguistic, the learning rate as LR, sequence length as SeqLen, and the number of trainable parameters as # Params.

Metric We propose CEEU metric to verify the situation switching function of the CD module. The conversation module generate the response based on dialogue history as shown in Figure 2. We input an utterance of out-of-situation in a random conversation turn in dialogue history, and measure whether the CD module detected it correctly.

Details of the corresponding procedures are described in Algorithm 1. SITUATION-CHAT denoted as D, contains a dialogue history $H_{n,t}$ for each situation n and a given turn t. Specifically, $H_{i,t}$ comprises gold answer $utt_{i,t}$, and utterance candidates $utt_{j,t}$ extracted from the different situation j. Once CD module correctly predicts the inclusion of the gold answer in a dialogue history, it is classified as a true positive (tp); otherwise, as a false negative (fn). Similarly, dialogue histories containing distractors are also classified into true negative (tn) or false positive (fp), based on the model judgment. Eventually, the accuracy is estimated based on the entire test set, where the number of utterances classified to the positive label and the negative label is set to be the same ratio.

D.2 Results

Table 9 presents the experimental results of the CD module, which are evaluated by the CEEU metric for situation similarity and the CoLA dataset for linguistic acceptability. We fine-tune four pre-trained language models, including BERT, ALBERT, RoBERTa, and XLNet. The highest CEEU score is achieved by XLNet with a score of 0.628. The highest linguistic acceptability is achieved by BERT and XLNet with CoLA validation scores of 0.812 and 0.851 and CoLA test scores of 0.820 and 0.870, respectively.

Algorithm 1 CEEU

```
Require: $CD_{Module}$ = {MRPC}
$S_n \leftarrow \{S_1, \dots, S_n \mid 1 \le n \le |D|\}$
/* $S_n$ is the situation. */
/* $|D|$ is the number of dialogues. */
$t \in \{1, 2, 3\}$
$H_{n,t} \leftarrow \{utt_{n,1}, utt_{n,2}, \dots, utt_{n,t} \mid 1 \le n \le |D|\}$
/* $utt_t$ is an utterance at turn $t$ of $n$th dialogue. */
procedure CEEU($S_n, H_{n,t}$)
    for $i \leftarrow 1, n$ do
        $H_{i,t} \leftarrow H_{i,t-1} + \left\{ \begin{array}{c} utt_{i,t} \\ \text{or} \\ utt_{j,t} \text{ where } i \neq j \end{array} \right.$
/* $utt_{j,t}$ is randomly chosen from other dialogues. */
        ContextScore $\leftarrow CD_{Module}(S_i, H_{i,t})$
        if $utt_{i,t}$ in $H_{i,t}$ then
            if ContextScore $\geq 50$ then
                $tp \leftarrow tp + 1$
            else
                $fn \leftarrow fn + 1$
        else if $utt_{j,t}$ in $H_{i,t}$ then
            if ContextScore $\geq 50$ then
                $tn \leftarrow tn + 1$
            else
                $fp \leftarrow fp + 1$
    return $Acc \leftarrow \frac{tp + tn}{tp + tn + fp + fn}$
```

Module	Model	CEEU	CoLA		GLUE	BLEU
			Valid	Test		
CD	BERT	0.476	0.812	0.820	-	-
	ALBERT	0.484	0.728	0.736	-	-
	RoBERTa	0.623	0.739	0.755	-	-
	XLNet	0.628	0.851	0.870	-	-
GEC	Park et al. (2020)	-	-	-	58.12	73.82

Table 9: Experimental results of CD module and GEC module. CEEU is used to evaluate situation similarity. We measure linguistic acceptability using valid and test set of CoLA. The performance of the GEC module is evaluated based on GLUE and BLEU.

D.3 Verification of GEC Module

The experimental result of GEC module is presented in Table 9. The performance of the GEC module is evaluated based on the GLUE (Napoles et al., 2015) and BLEU (Papineni et al., 2002) score. Further, the NS2DS (Park et al., 2020) is used as the experimental model, and Park et al. (2020) is

Dimension	4	3	2	1
Accuracy	Uses sentence structure, vocabulary, and grammar correctly without errors	Uses sentence structure, vocabulary, and grammar correctly with few errors	Uses sentence structure, vocabulary, and grammar correctly with several errors	Uses sentence structure, vocabulary, and grammar correctly (many errors)
Communication	Communicates thoughts and be understood without errors	Communicates thoughts and be understood with few errors.	Communicates thoughts and be understood with several errors.	Not able to communicate thoughts or be understood
Fluency	Communicates clearly and smoothly	Communicates clearly and smoothly with a little hesitation	Able to communicate with some prompts	Not able to communicate clearly or smoothly
Comprehension	Understands and always responds appropriately	Understands most verbal cues and mostly responding appropriately	Understands some verbal cues and sometimes requiring prompts	Not able to understand verbal cues or to respond
Maturity of the language	Includes details beyond the minimum requirements (word-choice/expressions/ gestures)	Includes details beyond the minimum requirements	Includes minimal or no details beyond the minimum requirements	Not able to utilize the language well

Table 10: Five dimensions of English-speaking performance evaluation.

referred to for the experimental results.

E Human Evaluation Measures

We ask the following additional questions to the evaluators to assess the quality of the model-generated utterances.

- **Fluency**: Fluency of the dialogue on a scale from 1 to 5, where 1 is “not fluent at all”, 5 is “extremely fluent”, and 3 is “fine”.
- **Situation Consistency**: Situation consistency on a scale of from 1 to 5, where 1 is “not maintained at all”, 5 is “well maintained”, and 3 is “fine”.
- **Engagingness**: Engagingness disregarding fluency from 1 to 5, where 1 is “not engaging at all”, 5 is “extremely engaging”, and 3 is “fine”.

F Details of User Evaluation

F.1 English-speaking Performance

To measure English-speaking performance, we recruited two English teachers who graduated with degrees in English education as evaluators. Following Chien et al. (2020), we modified the methodology to measure five dimensions: Accuracy, communication, fluency, comprehension, and maturity of language. Each dimension was evaluated on a scale of 1 to 4, with 1 being the lowest and 4 being the highest. Table 10 provides a description of each dimension. The inter-rater reliability for the evaluators on the English speaking ability had a cohen’s kappa value of 0.58 ($p < 0.001$).

F.2 English Grammar Test

For the Grammar evaluation, we use multiple-choice questions to evaluate the participants’ knowledge of English grammar rules. The questions cover a range of topics, such as verb tenses, prepositions, and articles. We administer a grammar test consisting of 20 items for both the pre-test and post-test.

F.3 English Learning Anxiety Scale

The measure for English learning anxiety, which consists of 33 items, was adapted from the Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986) with a 5-point Likert-type rating, such as “I never feel quite sure of myself when I am speaking in my foreign language class.” A higher rating indicates higher English learning anxiety. The Cronbach’s alpha of the measure was 0.95.

G PEEP-Talk Satisfaction Surveys

Question
Are you satisfied with the overall performance of PEEP-Talk?
Do you think situational dialogue is helpful for conversational learning?
Are you satisfied with the result of the context detector module?
Are you satisfied with the performance of the grammar error correction module?
Do you think the chatbot’s conversation is natural and contextual?

Table 11: Questions of PEEP-Talk satisfaction survey.

We conduct user satisfaction surveys on our platform. Five questions are selected to ask users to evaluate on a Likert scale(Likert, 1932) of “satisfied”, “neither agree nor disagree”, and “dissatisfied”. The survey questionnaires are listed in Table 11. We show the results of the survey in Figure 8.

As a result of the survey, most people express satisfaction with the performance of PEEP-Talk, and more than 70% of the respondents answer that the utterances of the chatbot are natural and contextual. Also, 70% to 80% of respondents answer that situational dialogue is educationally beneficial for conversational English, and grammar correction is helpful. However, the satisfaction of the CD module is 46%, which is relatively low. We speculate that this is due to the CD module switching topics, even if the user wants to talk more about a particular situation.

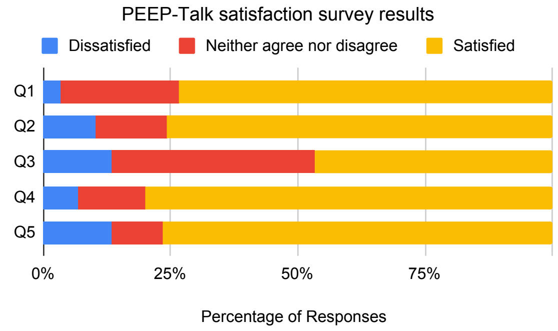


Figure 8: Results on the PEEP-Talk satisfaction survey

G.1 Response Comparison

Table 12 presents a qualitative analysis of the performance of PEEP-Talk in comparison with DailyDialog, Andy, and Speak Now. In each situation, PEEP-Talk provided responses that were contextually appropriate and reflected a natural conversation flow, indicating its potential to be an effective tool for real-world English education. In contrast, the responses generated by the other chatbots sometimes appeared to be unrelated to the topic or situation, suggesting that they may not be suitable for use in English education.

Situation	**PEEP-Talk**	**DailyDialog**	**Andy**	**Speak Now**
Bookstore There are so many new books this month.	I know, I am looking for a good Korean novel these days.	I know, I'm just impatient.	Do you like to travel?	I'm Okay.
Requesting packaging Can I get the leftovers to go?	You can, but we don't have a container for soup dishes.	I'm sorry, I don't have any leftovers.	Sure, will you write me later?	I'm looking at you
At the art museum Which way should I go to see the special exhibition hall?	If you go to the right, you'll see the entrance of the exhibition hall.	I think you have to go to the main hall.	Is it a good place?	Oh, You hurt my feelings
Customer service I ordered sweet potatoes to go on a diet, but there were some rotten ones included, can I get them exchanged?	I'm sorry, but you can't exchange or refund food that has already been sent.	I don't think you can exchange them.	My name is Andy. What is your name?	I'm okay.
Returning I bought a cell phone case and it's not pretty so I want to get a refund.	I am sorry but cell phone cases are non-refundable.	I bought a case and it's not pretty so I want to get a refund. english haiku bot	Got it	Oh, You hurt my feelings
Find product I'm looking for a lotion for dry skin, is there a cosmetic that is not too sticky?	If you want to use it as a base, I recommend this lotion for oily skin.	I use moisturizer.	I see	Glad to hear that.
Travel agency How much is a cheap flight from Incheon to New York airport?	I heard it is about 80 dollars if you depart from here.	I think it's about the same price as a flight to the US from the UK.	Got it	Let's keep going.
Clothing store I want to buy these clothes, can I try it on first?	You can try it on in the change room, were you not able to find it?	I don't think you can buy them, but you can buy the shirt.	To buy these clothes is a great goal.	What color is your hair?
Job interview Please tell me about your previous job experiences if any.	I worked in a bookstore as a part-time job and had a lot of free time.	I have a few. I've been in the service for a few years.	As you know my name is Andy and I want to help you in learning English.	Let's keep going.
IT/computer Is this the IT department? The program isn't working, please check it.	Yes, which program isn't working?	I'm not sure what you mean.	Thank you! I'm so happy to hear that, I love talking to you too	I'm Okay.
Daily Conversation It seems like you can't go home often ever since you moved out from home.	I can't go home often because I have too many friends who are also students.	I've been home for a few months now.	Got it	Thanks. Have a nice day.
Inquiring about prices Are there clothes on sale now?	Yes, we have sales clothes on this side separately.	I don't think so. I think they are just a limited time thing.	I see	Great. Man.
Duty-free shop I was told that the cigarettes are really cheap if bought at duty-free shops.	I heard that too, but isn't it just cigarettes?	I've never seen a duty free shop that has them.	Okay	I don't understanding what you are saying.

Table 12: Comparison of the utterances generated by PEEP-Talk, DailyDialog Andy, and Speak Now.

H Example Conversations

We show a sample conversations of Andy, Speak Now, and PEEP-Talk.

H.1 Conversations between Humans and PEEP-Talk (Ours)

PEEP-Talk (Ours) Conversation 1	
Situation:	**At the cosmetic store**
Human:	*What are some products on sale today?*
PEEP-Talk:	*We also have a 10 percent discount on lotion and toner.*
Human:	*This is the product I am using, and do you still sell it?*
PEEP-Talk:	*We do, but we only have a small amount left.*
Human:	*I'd like to buy a set of cosmetics.*
PEEP-Talk:	*I'll go get it for you.*
Human:	*Thank you. Have a nice day*

PEEP-Talk (Ours) Conversation 2	
Situation:	**Evaluating food taste**
Human:	*I heard that salmon salad is really famous at this restaurant.*
PEEP-Talk:	*Really? I love salmon salad . I also heard that the Stir-Fried mushrooms are also very famous*
Human:	*Then, shall we order a salmon salad and a pizza and share them?*
PEEP-Talk:	*I don't think I can eat Pizza alone. Let's order some more and share them.*
Human:	*Today the weather is good.*
PEEP-Talk:	*You are right. The temperature is good and the wind is good.*
Human:	*Good.*
PEEP-Talk:	`--Situation switch required--`
PEEP-Talk:	*Why don't we speak another situation? please, click the different situation!*

PEEP-Talk (Ours) Conversation 3	
Situation:	**Situations diagnosed in the hospital**
Human:	*I have a headache*
PEEP-Talk:	*When did the headache start?*
Human:	*I started having symptoms last week.*
PEEP-Talk:	*It seems you have a persistent headache.*
Human:	*And I have chills and fever.*
PEEP-Talk:	*I think we need to do a detailed examination.*
Human:	*Yes, thank you. Where is it going to take a place?*
PEEP-Talk:	*I will go to the central nervous system and tell you the location.*
Human:	*Okay. Thank you*
PEEP-Talk:	*Here is a paper with a magnet attached to it . please follow me.*

PEEP-Talk (Ours) Conversation 4	
Situation:	**School life**
Human:	*Professor, when will you hand out the midterm results we took last week?*
PEEP-Talk:	*I am writing them now . do you want them by any chance?*
Human:	*Could you please tell us the highest and lowest scores in advance?*
PEEP-Talk:	*The students usually get an 85 on the Midterms and the lowest is 70 points.*
Human:	*Okay. Thank you*
PEEP-Talk:	*You're welcome . Let me know if you have any questions.*
Human:	*How do we decide the groups? Will you be randomly assigning them?*
PEEP-Talk:	*I will assign them based on how many friends are in the class.*
Human:	*How many people are assigned per group?*
PEEP-Talk:	*I will assign 5 people to a group of 15.*

Table 13: Example conversations between humans and our proposed PEEP-Talk. It is able to not only converse in a certain real-world situation but also can switch the situation or topic when the users input the utterance predicted as it is out of the situation (Conversation 2).

H.2 Conversations between Humans and Andy

	Andy Conversation 1
Andy:	*Hello.*
Human:	*Good morning.*
Andy:	*My name is Andy. What is your name?*
Human:	*My name is Jun.*
Andy:	*That's great! I am happy for you.*
Human:	*What are you doing?*
Andy:	*I was doing some exercises. I always feel sleepy and sad without physical activity.*
Human:	*Me too. I have to work out in the morning and evening.*
Andy:	*Interesting.*

	Andy Conversation 2
Andy:	*Where are you from?*
Human:	*I came from Canada.*
Andy:	*Is it a good place?*
Human:	*I like a fencing.*
Andy:	*Oh, I want to visit it now*
Human:	*Come here. Anytime.*
Andy:	*Nice. Do you have brothers or sisters?*
Human:	*Yes.*
Andy:	*And I have a little sister. She studies modern art in Rome. She is very creative person. I'm sure she will become a famous artist.*

	Andy Conversation 3
Andy:	*Hi, Jun. What's up*
Human:	*How are you?*
Andy:	*I'm doing great, thank you my friend.*
Human:	*What do you think of the Lakers this season?*
Andy:	*It's very sunny!*
Human:	*This team is too old. We need new blood.*
Andy:	*Yes, I'm completely fool*
Human:	*All our family members always have been supporting the same team.*
Andy:	*Nice. Do you know any jokes?*
Andy:	*No*

	Andy Conversation 4
Andy:	*How do you think cities will change in the future?*
Human:	*Nothing much, just thinking about the good old days when I was a kid. How about you?*
Andy:	*I'm doing great, thank you!*
Human:	*That's great to hear! I hope you have a great rest of your day! :)*
Andy:	*I understand. Of all the people in history that have reached 65 years of age, half of them are living right now.*
Human:	*Wow, that's a lot of people to live with. Do you have any siblings?*
Andy:	*Yeah, amazing. What's your father like?*
Human:	*He's a great man. He's always been there for me when I needed him the most.*
Andy:	*Got it. Can you tell about your parents?*

Table 14: Example conversations between humans and Andy. The responses of Andy are restricted and cannot cover a variety of topics.

H.3 Conversations between Humans and Speak Now

Speak Now Conversation 1	
Speak Now:	*What's up. Did you have breakfast?*
Human:	*No, I haven't yet.*
Speak Now:	*Oh, I'm sorry to hear that. Do you like seafood.*
Human:	*Yes. I like shrimp.*
Speak Now:	*How we should go to Busan someday to eat seafood. Do you like meat?*
Human:	*Yes, I like it.*
Speak Now:	*Good. We should eat meat someday. What's your favorite food?*
Human:	*My favorite food is sushi*
Speak Now:	*How we have same taste. Do you like breads?*

Speak Now Conversation 2	
Speak Now:	*What's up. Did you have breakfast?*
Human:	*Yes, I ate 15min ago.*
Speak Now:	*I'm doing okay man. What would you like to talk about today?*
Human:	*Let's talk about cars*
Speak Now:	*No. I'm not sure about that. We Should talk about something else.*
Human:	*Why don't we talk about colors?*
Speak Now:	*No. I can not understand. Have you ever went study aboard?*
Human:	*No. I have not. But, I want to go further day.*
Speak Now:	*Oh. Should Have you ever been to England?*

Speak Now Conversation 3	
Speak Now:	*Did you Sleep well, last night?*
Human:	*I didn't sleep very well*
Speak Now:	*Oh. I sorry to hear that. Do you like drinking alone?*
Human:	*No, I did not like alcohol*
Speak Now:	*What's your favorite alcoholic drink?*
Human:	*I did not like alcohol*
Speak Now:	*Have you invested in Bitcoin before?*
Human:	*No, I have not*
Speak Now:	*A lot of people lost money for that. What would you like to talk about today?*
Human:	*I want to talk about soccer,*
Speak Now:	*I don't like topic. Let's try different one*

Speak Now Conversation 4	
Speak Now:	*Hey Dude did you sleep well last night*
Human:	*I did sleep well*
Speak Now:	*Do you like your Boss?*
Human:	*No I don't like my boss*
Speak Now:	*I understand you I think most people do. Are you satisfied your job?*
Human:	*No I do not satisfy my job .*
Speak Now:	*I understand you it can always be fun. Which country are you from?*
Human:	*I came from Korea*
Speak Now:	*How you live in a good place Where do you live?*

Table 15: Example conversations between humans and Speak Now. We used free conversation mode of Speak Now. As Speak Now uses hand-crafted rules and pattern matching when making the utterances, the utterances are limited to its own pre-built templates.