IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 June 2023, accepted 15 July 2023, date of publication 21 July 2023, date of current version 27 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297652

RESEARCH ARTICLE

Enhancing Conversational Model With Deep Reinforcement Learning and Adversarial Learning

QUOC-DAI LUONG TRAN[1], ANH-CUONG LE[1], AND VAN-NAM HUYNH[2], (Member, IEEE)
[1]Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
[2]School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Japan

Corresponding authors: Anh-Cuong Le (leanhcuong@tdtu.edu.vn) and Van-Nam Huynh (huynh@jaist.ac.jp)

This work was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant 102.05-2020.26.

ABSTRACT This paper develops a Chatbot conversational model that is aimed to achieve two goals: 1) utilizing contextual information to generate accurate and relevant responses, and 2) implementing strategies to make conversations human-like. We propose a supervised learning approach for model development and make use of a dataset consisting of multi-turn conversations for model training. In particular, we first develop a module based on deep reinforcement learning to maximize the utilization of contextual information serving as insurance for accurate response generation. Then, we incorporate the response generation process into an adversarial learning framework so as to make the generated response more human-like. Using these two phases in combination eventually results in a unified model that generates semantically appropriate responses that are also expressed naturally as human-generated in the conversation. We conducted various experiments and obtained a significant improvement compared to the baseline and other related studies.

INDEX TERMS BERT, chatbot, reinforcement learning, sequence to sequence, generative adversarial nets.

I. INTRODUCTION

Chatbot research focuses on creating intelligent agents for human-machine interaction in a fashion similar to human-to-human communication. Chatbots have a variety of uses, including customer service, e-commerce, healthcare, finance, education, entertainment, HR, and emergency services. They can assist customers with inquiries and provide information and support in account management, scheduling, medical advice, and recruitment. As a result, the field of artificial intelligence and natural language processing is witnessing more research on building Chatbot applications. Recently, there have also been breakthroughs in the field, leading to practical applications such as ChatGPT.

The evaluation of a Chatbot model is based on its ability to engage in natural, human-like conversations. In 1950, Alan Turing introduced the Turing test as the standard measure of a chatbot's ability to think like a human. The test, which is used to determine if a chatbot can imitate human behavior, requires that a human judge be unable to reliably distinguish the chatbot's responses from those of an actual human.

It is clear that a chatbot can be considered to have human-like quality if it can generate responses that are relevant, meaningful, and natural in conversations with humans. In an attempt to achieve that goal, we have recently developed a new model for conversational agents inspired by the Turing test and the idea of the adversarial learning method [1]. Particularly, we designed a model based on deep neural networks that allow generating accurate responses optimized by the mechanics of imitating human-generated conversations, which experimentally results in improvement of the model performance. However, the model cannot evaluate the effects of contextual information in an entire conversation, such as the relationship of contexts and their influence on future outcomes. This study aims to overcome this disadvantage of our previously developed model by analyzing various contextual information that affects utterances in a conversation so as to incorporate them into development of the model to improve the coherence and consistency of a multi-turn conversation.

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

As observed, the majority of current chatbot systems can be classified into two categories: pattern-based and learning-based approaches. A rule-based method [2] is one of the earliest pattern matching methods for building chatbots. In this method, a chatbot is trained based on some predefined rules to answer questions. Upon receiving an utterance as input, the chatbot will match it with the most similar question pattern in a predefined set of rules and fetch the corresponding answer as a response. The chatbot is ineffective in this approach when input does not match the predefined rule sets. Furthermore, most rule-based systems build predefined rules manually, making it difficult for the systems to extend their conversational knowledge.

Another method in the pattern-based approach is a kind of information retrieval (IR) [3], [4], [5], which is useful when there is a large set of dialogues. In this method, the chatbot is trained on a set of questions and corresponding answers by matching the user input to a set of questions from a dataset and selecting a corresponding answer using pattern-matching algorithms. If IR-based systems are trained on a large enough dataset, they can achieve decent performance due to the expert domain knowledge they contain. However, they still cannot generate new answers as they rely only on the database of existing responses. They also have difficulty in understanding the semantic difference between different input contexts, making them suitable only for single-turn communication but not for forming human-like conversations.

Recently, chatbot models have been developed using machine learning-based methods, where they learn from a set of conversations in the training data. These models, such as LSTM or Transformer, usually have two components: an encoder to encode the information in the user's question and a decoder to generate the chatbot's response. Chatbot systems developed using machine learning methods have been able to generalize the learned data in deep learning models, allowing them to answer new questions and generate dynamic and adaptive responses, rather than being limited to pre-determined patterns as in pattern-based models.

Despite some advantages of machine learning-based approaches, these systems currently face some limitations in generating responses. The main reason is that most of these models are to make use of only the current utterance of the conversation to generate the corresponding response (i.e the next utterance). This results in not having the necessary context for the model to generate appropriate responses. And so in the case of conversations where similar answers to different questions exist, it makes the system more confusing. For example, because there are many generic responses in the training dataset then a Seq2Seq based model tends to generate generic and dull responses such as ''I don't know'' regardless of the input utterance [6], [7], [8]. Such kind of responses can make the conversation stop or fall into an infinite loop after three turns [8].

Notably, large language models such as GPT-3 (Generative Pre-trained Transformer 3) have been recently developed to excel in language comprehension in context, enabling them to exhibit exceptional performance on many NLP tasks [9]. In particular, GPT-3 that is an autoregressive language model has been released by OpenAI in 2020 and then further improved into GPT-3.5, which was used to build the so-called ChatGPT, the hottest AI chatbot at the moment. While systems like ChatGPT have leveraged large language models trained on extensive datasets using supervised fine-tuning (SFT) and reinforcement learning (RL) from human feedback (RLHF) techniques [10], their strength lies in significant model size and wealth of knowledge that, therefore, requires extremely high computational cost and resources and it seems only tech giants can afford.

In low-resource scenarios, our research is specifically focused on developing models that can effectively operate with limited data and low computational resources. To this end, besides the current utterance, our proposed model will make use of additional information from the context of the current utterance when generating a response. In practice, humans always use contextual information in their decision-making. There have also been other studies that used additional previous utterances as the context to the left of the current utterance in the model to generate responses, as in studies [6], [11], [12]. Our proposal has the advantage and is more general than these studies, as we will use both the left and right contexts of the current utterance. Especially the right context will be used only in the model training phase, and we will use the RL strategy for this purpose. For a chatbot to be more like a human, it needs to generate suitable responses and ensure that the responses have a clear goal and fit with the conversation context.

Inspired by the advantages of GANs [1], in this work we apply the adversarial learning strategy of GAN to fine-tune the generation model of RL for generating responses like a human. Moreover, because the RL model uses a lot of context-based information in a conversation for response generation, it is possible to experience overfitting, and so the use of GAN also helps to reduce this phenomenon. The discriminator component in the GAN will assess whether a generated response is genuine or synthetic (i.e. human-like or artificial). We use this evaluation score as an additional factor in constructing the reward function for the RL algorithm in the proposed model.

In summary, the main contributions of this study are two-folds:

- We propose a response generation model which takes into consideration of both the left and right contexts of the current utterance to generate relevant responses in the conversation. This idea is implemented by using a deep RL model.
- We utilize adversarial learning techniques to enhance the naturalness of responses generated by a newly developed model based on a hybrid reinforcement and adversarial learning approach, in which the adversarial module's discriminator acts as a human evaluator in the Turing test, and feedback from this evaluator is applied as a

reward for RL to train the chatbot towards human-like behavior.

Overall, this study aims to develop a model that focuses on specific scenarios or predefined scopes, capable of handling conversations with desired goals. Our goal is to enhance the coherence of the chatbot's future responses by taking into account the conversation history. This approach addresses the drawback of generating irrelevant answers, often resulting in dialogue breakdown in traditional models. Ultimately, it aims to improve the overall user experience by enabling the chatbot to provide more relevant and contextually appropriate responses.

The rest of this paper is structured as follows: Section II briefly summarizes previous research on conversational agents. Section III describes the problem formulation, the backbone framework concept and model development. Section IV presents our experimental results and analysis. The paper is concluded in Section V.

II. LITERATURE REVIEW

This section reviews previous studies on conversational agents, which have mainly focused on training chatbot models based on two main approaches: pattern matching and machine learning approaches. We will also discuss their strengths and weakness; and the reason why they are still incapable of passing Turing's test.

Initially, inspired by the Turing test, several chatbot systems were developed with the goal of passing it. One of the first such systems is ELIZA [13], developed by Joseph Weizenbaum in 1960s at the MIT Artificial Intelligence Laboratory. ELIZA is a rule-based chatbot that uses pre-defined rules to imitate conversation through pattern matching and substitution. The program responds to users by identifying keywords in the input message and transforming it into a response using the corresponding rule. Another chatbot, PARRY, is an extension of ELIZA [14] developed at Stanford. PARRY introduced novel improvements compared to ELIZA. Unlike ELIZA, PARRY generated responses not only based on the previous sentence but also kept track of the user's emotional state by simulating a schizophrenia patient. During 1995 to 2000, Artificial Intelligence Markup Language (AIML) was developed to build Knowledge Bases for chatbots using the Pattern Matching approach. AIML is an open-source XML-based language. ALICE was the first chatbot with a Knowledge Base built using AIML [2].

ALICE's Knowledge Base, which consisted of around 41,000 templates, was much larger than ELIZA's 200 rules. However, ALICE still lacked the intelligence to produce human-like responses. The weakness of pattern-matching approaches is that they often result in repetitive automatic responses and lack the spontaneous nature of human responses. Pattern-matching-based chatbots simply select answers from a database and only take into account the previous message without considering the context of the conversation. To make a chatbot more human-like, it must generate appropriate responses that have a clear purpose and align with the context of the conversation.

In contrast to pattern-matching approaches, chatbots based on machine learning use NLP techniques to derive knowledge from input or learn the context of a conversation. They not only consider the current message but also take into account the context of the conversation. They are not limited to providing a pre-defined response based on a rule for each user input. The core concept of machine learning is to train a model that maps input to output based on a set of input-output pairs in a dataset, learning from them.

A conversational system that uses NLP techniques was created with Support Vector Machine (SVM) and Logistic Regression classifiers [15]. The results of comparing this model to rule-based and retrieval-based models showed significant improvement in naturalness. Many recent studies have used RNN-based models to build chatbots. The study [16] utilized an RNN model that incorporates previous context in a dialogue by feeding user input and prior information directly to outputs through a series of weights. This enables the model to transfer conversation history knowledge to current responses. LSTMs, an extension of RNN, can enhance chatbots by allowing them to refer to prior information and learn long-term dependencies, as shown in [17]. The study [18] further improved context retention and knowledge utilization by combining RNN and bidirectional GRUs with an attention mechanism, resulting in better performance compared to the baseline, as shown in experiments with the Wizard-of-Wikipedia dataset.

Many studies based on the machine learning approach mainly employ a sequence-to-sequence (Seq2Seq) model. The Seq2Seq is about training models to convert a source from one domain to a target in another domain. The conversational agent can be formulated as a source-to-target task, in which the source is the user's message, and the target is the chatbot's response. A vanilla Seq2Seq model generates a target response based on the source message [19]. In [20], they used Seq2Seq (LSTMs) to build a deep-learning chatbot for the Portuguese language. However, their results cannot achieve high performance due to the need for low-resource language. The lack of training data is also a disadvantage of machine learning approaches. So, an explicit drawback is that the Seq2Seq models need an extensive training set for the training process. The MarkBot framework was also built and developed with the LSTMs structure [21]. Many recent studies showed that Seq2Seq-based chatbots tend to generate short and dull responses such as ''OKay'' and ''I don't know'' [22]. One of the weaknesses of the generative models is that the outputs can lack consistency [23]. These problems cause users to quickly realize the machine's presence in a dialogue. In order to resolve the problem, [24] suggested an evaluation metric to estimate the humanness of the chatbot. The metric captures characteristics of a human-like in conversation.

To address the limitations in multi-turn conversation models, a new approach considers conversation as a

decision-making process (DMP) with states, actions, and strategies. A conversation can be defined as a Markov Decision Process (MDP) or a Partially Observable Markov Decision Process (POMDP) [25], [26], [27]. As a result, RL techniques can be utilized to control state transitions, generate suitable actions (conversation utterances), and gather information from the user [28]. Chen et al. [29] proposed a DRL framework that implements a structured actor-critic model, which is trained in parallel on a dataset collected from various dialogue tasks. The model was evaluated on 18 tasks in PyDial and showed efficient and stable learning. Unlike most methods that use RL through trial-and-error learning, Offline RL [30] proposes using static datasets to train dialogue agents. An another study proposed an RL-based method to build a chatbot using a generation model that generates sequences for a task-oriented model [31]. The experiments demonstrate that this approach leads to more realistic conversations that better achieve task goals. A study designs dialogue systems to track the history of conversation [32], [33]. This information informs better decision-making for the next action or response selection and represents the user's goals while incorporating the dialogue history.

In recent years, GAN has became a highly popular generative model that demonstrates remarkable improvements in generation tasks. Initially proposed for image generation [34], GAN has been expanded to include discrete and textual data [35], [36], [37]. For example, a GAN-based model was used to generate poems, music, and speeches of President Barack Obama [38], showing significant improvement over baselines. The model was also applied in text-generation tasks. Additionally, an evaluation model based on GAN was first introduced to assess the quality and effectiveness of generated responses [39], [40], reducing the reliance on human evaluation [41].

As previously aforementioned, while AI chatbot based on the GPT architecture like ChatGPT has gained significant attention due to its great capability of generating human-like text in a wide range of domains and styles, it typically requires large data and computing resources to train and operate effectively. In low-resource scenarios, there remains a demand for solutions to effectively develop models capable of operating with limited data and computing resources.

In this study, we focus on exploiting contextual information in muti-turn conversations using RL. We build our model using GANs along with RL to achieve multiple goals, including avoiding overfitting by incorporating multiple context constraints in RL and training the system to generate human-like responses through adversarial learning.

III. THE PROPOSED MODEL

In this study, we address the problem in a multi-turn scenario using a dataset of conversations. Each conversation consists of a sequence of multi-turn dialogues $s_0, s_1, s_2, \ldots, s_{n-1}, s_n$, where s_t and s_{t+1} represent successive turns in the conversation between two agents. Building a chatbot model is viewed as a source-to-target mapping task, where given a source utterance x, the model learns mapping rules from training data to produce a corresponding output. We split the training dataset into n pairs $(\overline{s_t}, \overline{s_{t+1}})_{t=1}^{n}$, where each pair (s_t, s_{t+1}) is a consecutive turn in the dialogue. In this system, each user utterance $s_t = w_1^t, w_2^t, \ldots, w_{|s_t|}^t$ is paired with its corresponding output $s_{t+1} = \{w_1^{t+1}, w_2^{t+1}, \ldots, w_{|s_{t+1}|}^{t+1}\}$ to be predicted, where w_k^t is the k^{th} word in utterance s_t.

A. CONTEXTUAL CHATBOT WITH REINFORCEMENT LEARNING

Following the success of sequence-to-sequence models, conversational systems are trained end-to-end using the Maximum Likelihood Estimation (MLE) objective function. They have demonstrated impressive performance in response-generation tasks. While Seq2Seq models are suitable for input-output mapping problems, they still have several drawbacks for multi-turn conversations, including a lack of specificity in modeling dialogs and a lack of contextual information in the decoding process. These models are trained to generate the following response based on a previous turn. As a result, these models often generate generic or uninteresting responses like ''I don't know'' because there are many generic utterances in the training data. The decoder also fails to consider the full context of the conversation as it only generates responses based on the previous turn.

We propose to solve the above problems by formulating conversation as a RL problem and optimizing response generation using long-term rewards. In recent years, RL has gained significant attention in chatbot development. This technique plays an important role in the success of ChatGPT. In this work we also utilize RL to train the model for providing accurate responses. However, unlike GPT-based chatbots, which rely on a single response for computing reward scores, our model considers multiple responses to ensure a better understanding of the conversation's context. By considering a sequence of responses, our model can capture the flow and coherence of the conversation more effectively.

In a multi-turn scenario, we frame response generation as a source-to-target problem, treating conversation history as the source and the next response as the target. The proposed model leverages the long-context success of a conversation by generating an utterance that is conditioned on the impact of the generated response in an ongoing dialogue. To do this, we first predict the best response for the historical context, then fine-tune the model based on the desired outcome and future context.

To achieve this goal, inspired by [8], we design a simulated conversation by having two chatbots communicate with each other (as shown in Fig. 1). The Seq2Seq pre-training enables the model to generate responses consistent with the conversation history, while the RL optimizes the responses for long-term goals. The simulation works as follows: at the first step, the first agent takes an input sentence with the conversation history as contextual information c_L from the training

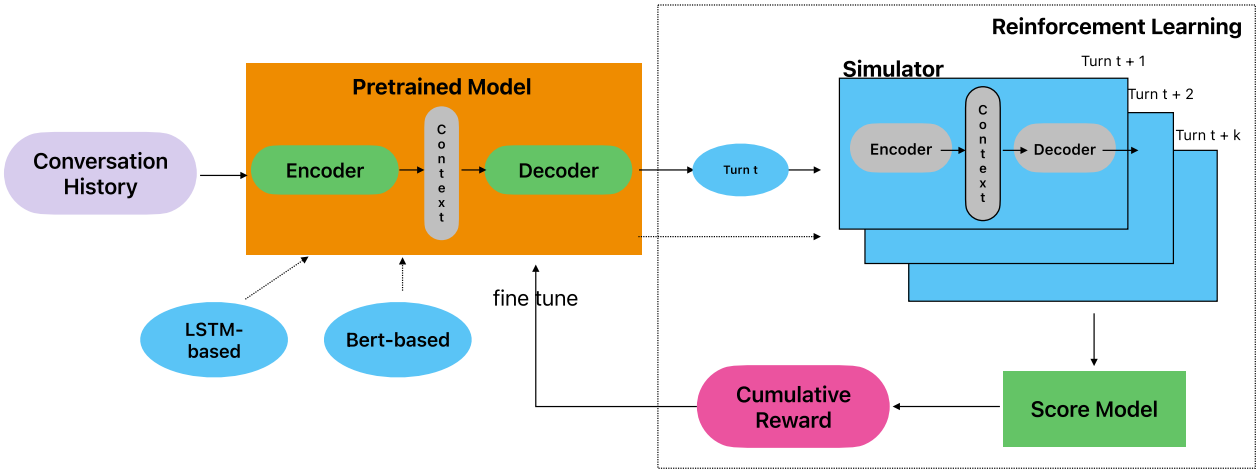

FIGURE 1. Contextual Deep Encoder-Decoder with Reinforcement Learning.

dataset and encodes it into a vector representation. The agent then decodes the vector representation to generate a response p for the next turn. The second agent updates the simulation state by combining the conversation history with the output p. It immediately encodes this new state into a vector representation and decodes it into a new response, which is fed back to the first agent and repeated.

Let (S, A, P, R) represent the MDP as a framing of the problem of learning from interaction to achieve a goal [28], where S is the set of states defined as the dialogue history, and A is the set of actions as generated responses. The policy P is the transition probability distribution π, implemented as a pre-trained encoder-decoder model. The policy model π is initialized using a pre-trained Encoder-Decoder model:

$$\pi = p_{Encoder\_Decoder}(a|[m, c_L]) \tag{1}$$

Given an input m and its context c_L, we generate a list of candidate responses A as follows:

$$A = \{a|a \sim \pi\} \tag{2}$$

The reward function R is a forward-looking metric that assigns a reward to each action of taking a as representing the desired outcome of the conversation. It is crucial in maintaining the coherence and effectiveness of the chatbot. The model learns to maximize the reward function to achieve the desired goals.

The Maximum Likelihood Estimation (MLE) objective function has limitations in capturing the goal of the conversation and can result in inconsistencies in the generated content [22], [23]. To address this, we use the virtual conversation context c_R to exploit the relationship between consecutive turns, and define the rewards to ensure that the generated responses are relevant to the conversation context.

Let h_t and h_{t+1} be two representations of consecutive responses s_t and s_{t+1} generated by the system at time t and $t + 1$, respectively. The coherence of the dialogue is maintained based on the cosine similarity between the two representations. The reward for each action of taking a_t as the response is computed as follows:

$$r_1(a_t) = cos(h_t, h_{t+1}) = cos(\frac{h_t.h_{t+1}}{\|h_t\|\|h_{t+1}\|}) \tag{3}$$

This score assesses the suitability of the generated responses by avoiding scenarios where the responses have a high probability of being generated but are not relevant to the history of the conversation. To guarantee their relevance, we use the mutual information between the current turn and the previous conversation to ensure their appropriateness.

In addition to the first reward, we propose a second reward to encourage the chatbot to contribute new information at each turn. This second reward can capture the relationships within the sentence and detect subtle changes in the content, thus providing new information for the conversation and maintaining its coherence and momentum. This can keep the conversation active and continuing without losing the conversation's coherence.

$$r_2(a_t) = \frac{1}{N_{a_t}} \sum_{w_p \in a_t} \min_{w_q \in s_{t+1}} \left\{ cos\left(\frac{w_p.w_q}{\|w_p\|\|w_q\|}\right) \right\} \tag{4}$$

where N_{a_t} is the number of tokens in the utterance a_t and, with an abuse of notation, w_p and w_q represent the embedding vectors of tokens in a_t and s_{t+1}, respectively. The final reward at step t for a given state s and action a is calculated as a weighted sum of the component rewards, as follows:

$$R(a_t) = \lambda_1 r_1(a_t) + \lambda_2 r_2(a_t) \tag{5}$$

where $\lambda_1 + \lambda_2 = 1$.

In RL, the objective of the agent is to maximize the expected reward from its actions [42]. The coefficients λ_1 and λ_2 are used to assign weights to each component reward, indicating their relative importance in the RL process. The condition $\lambda_1 + \lambda_2 = 1$ in this context is used to ensure that the weights are correctly normalized and their values lie

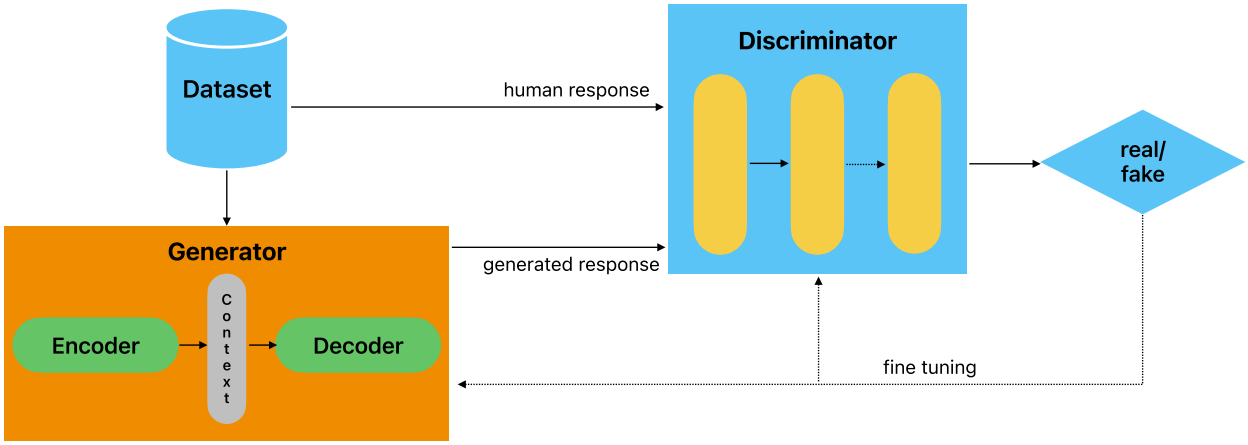

FIGURE 2. The illustration of GAN-based model.

within a valid range. The specific choice of weights λ_1 and λ_2 depends on the importance of each component reward, and the desired balance between them. The focus is to maximize the following objective function:

$$J(\theta) = \sum_{[a_t,\ldots,a_{t+k}]\in A} \pi_\theta(a_t,\ldots,a_{t+k})R(a_t,\ldots,a_{t+k}) \quad (6)$$

where a_t is an utterance in turn t and θ represents the set of parameters in the model, $R(a_t,\ldots,a_{t+k})$ is the cumulative reward associated with the sequence of utterances $a_t,\ldots,a_{t+k}$ defined as follows:

$$R(a_t,\ldots,a_{t+k}) = \sum_{\tau=t}^{t+k} \gamma^{\tau-t} R(a_\tau) \quad (7)$$

in which τ is a discount factor in the RL algorithm that determines the significance of future rewards compared to immediate rewards.

The training objective is to find the optimal parameters for the model that maximizes the expected reward. Note that the equation (6) can be rewritten as expectation:

$$J(\theta) = \mathbb{E}_{a_t,\ldots,a_{t+k}\sim\pi_\theta(a_t,\ldots,a_{t+k})}[R(a_t,\ldots,a_{t+k})] \quad (8)$$

The algorithm for modeling the contextual chatbot using RL is summarized in Algorithm 1.

Algorithm 1 RL-Chat

Require: Input sequence (X), ground-truth output sequence (Y) and history conversation (C_L)

1: Initialize the policy model π_θ
2: Setup the policy π with a pretrained model
3: **for** number of training iterations **do**
4: Run policy π and get response a_t
5: Run the simulator to get sequence of sentence $a_t,\ldots,a_{t+k}$, with $a_i \sim \pi_\theta$
6: Observe the sequence and calculate the reward according to (7)
7: Calculate the expected reward according to (8) and update the parameters of the model
8: **end for**

B. HUMAN-LIKE CHATBOT WITH ADVERSARIAL LEARNING

Such models as ChatGPT and Bard rely on large language models to enhance their capabilities of natural language generation. In the proposed approach, instead of constructing a large language model to achieve highly natural and human-like language generation, we employ adversarial learning techniques to make the model capable of generating natural language. As is well known, the characteristic feature of the GAN model is its discriminator component, which is used to steer the generator's output towards the desired direction. Here, we aim to design a discriminator that can distinguish between human-generated and model-generated utterances. By having the generator intentionally "fool" the discriminator, we will train the model to produce responses that are as natural as those made by humans.

Fig. 2 provides a representation of how this model operates. We borrowed the concept of the Turing test and the adversarial training technique [43], [44]. The proposed model trains two components simultaneously, a generator G and a discriminator D. The two networks are trained towards opposite objectives, as described below:

- The generator G defines the probability distribution over the dataset and generates a response y based on the dialogue history x. The goal of the generator is to deceive the discriminator into thinking that the output y is from a human.
- The discriminator D is a binary classifier that takes as input a sequence of utterances and outputs a label indicating whether the input was generated by a human or machine. The objective of the discriminator D is to distinguish between machine-generated and human-generated utterances.

The generator G is defined as a function $G(m, \theta_G)$ that takes a message m as input and uses θ_G as parameters. The discriminator D is defined as $D(x, \theta_D)$ and outputs a single scalar. It takes observed variables x as input and uses θ_D as parameters. The discriminator D is trained to distinguish between real data from the dataset and fake data generated by G. In this manner, the discriminator D functions as the evaluator in the Turing test.

Both the generator and discriminator have cost functions that are defined in terms of their parameters (θ_D and θ_G), and each network's cost function depends on the other network's parameters, but neither network can control the other's parameters. The cost function of the model is referred to as C. The generator is expected to minimize $C_G(\theta_D, \theta_G)$ while controlling only θ_G, and the discriminator is expected to minimize $C_D(\theta_D, \theta_G)$ while controlling only θ_D. We use G to maximize D's errors, and D is used to minimize its errors. G and D are trained separately. During training, G minimizes $\log(1 - D(G(x)))$. Formally the mini-max between G and D is given by [43]:

$$\min_{\theta_G} \max_{\theta_D} V(D, G) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(x))] \quad (9)$$

Algorithm 2 GAN-Chat

Require: The training dataset was split into n pairs $(s_t, s_{t+1})_{t=1}^{n}$ where (s_t, s_{t+1}) is the t^{th} pair of an input and its corresponding target

1: Initialize the Generator network g_θ, and the Discriminator network d_ϕ
2: Setup the Response Generator with pretrained Seq2Seq or BERT
3: **for** number of training iterations **do**
4: **for** k steps **do**
5: Get (real, fake) pairs data from dataset and Generator network;
6: Train Discriminator with (real, fake) pairs. (Perform gradient decent to update d_ϕ)
7: **end for**
8: Fine-tune Generator (Perform gradient decent to update g_θ)
9: **end for**

The algorithm is summarized in Algorithm 2. During training, the generator network is first trained based on the traditional Seq2Seq approach or a pre-trained BERT-based model. The generator takes a pair of human post and response from the dataset and uses the encoder to obtain a fixed-length context vector. The decoder then decodes this vector into a fake response. The human response from the dataset is selected as the real response for training the discriminator. The discriminator is trained to classify real responses drawn from the training set and fake responses generated by the generator. The generator strives to trick the discriminator while the discriminator's goal is to identify fake responses from the generator. In this way, the discriminator serves as the human evaluator in the Turing test.

C. COMBINATION OF REINFORCEMENT AND ADVERSARIAL LEARNING

The unique aspect of our proposed model is the combination of deep RL and GAN for generating both accurate and human-like responses, as graphically illustrated in Fig. 3. Based on deep RL, the generator learns to model the contextual information in the conversation by mimicking human talk. It takes into account historical dialogue through the left-context and considers the impact of future information flow on the dialogue through the right-context. The generator G is initialized using a policy model π as shown in (1).

We build upon the RL component as a backbone and integrate a discriminator from adversarial learning to evaluate the responses generated by the RL component. The output of the discriminator is used as an additional reward for the RL generator, encouraging the model to generate responses that are indistinguishable from human responses. That is, a new reward associated with the RL component is defined as follows:

$$r_3(a_t) = D(G(a_t)) \quad (10)$$

This reward acts as teacher feedback, providing insights into the model's contribution to achieving learning goals. As such, by incorporating this RL component into the model, the overall reward at the time t previously defined in (5) is now updated as follows:

$$R(a_t) = \lambda_1 r_1(a_t) + \lambda_2 r_2(a_t) + \lambda_3 r_3(a_t) \quad (11)$$

where λ_1, λ_2 and λ_3 are weights reflecting the relative contribution of each component reward to the overall one and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

The goal of RL is to maximize the sum of future rewards. By incorporating the factor $r_3(a_t)$ into the learning objective function of the RL model, there is a balance between generating a response that satisfies both the context requirements of the RL and the condition that the response belongs to the set of human-like responses. This also helps to reduce the overfitting in the mulitple contexts based RL model.

Similar as (8), we have the objective function for a virtual conversation as the sum of all responses:

$$\begin{aligned} J(\theta) &= \sum_{[a_t, \ldots, a_{t+k}] \in A} \pi_\theta(a_t, \ldots, a_{t+k}) R(a_t, \ldots, a_{t+k}) \\ &= \sum_{a_{t:t+k} \in A} \pi_\theta(a_{t:t+k}) R(a_{t:t+k}) \end{aligned} \quad (12)$$

The reinforcement algorithm maximizes the objective function using the gradient of it. The derivative of $J(\theta)$ is given as follows:

$$\partial_\theta J(\theta) = \sum_{a_{t:t+k} \in A} \partial_\theta \pi_\theta(a_{t:t+k}) R(a_{t:t+k}) \quad (13)$$

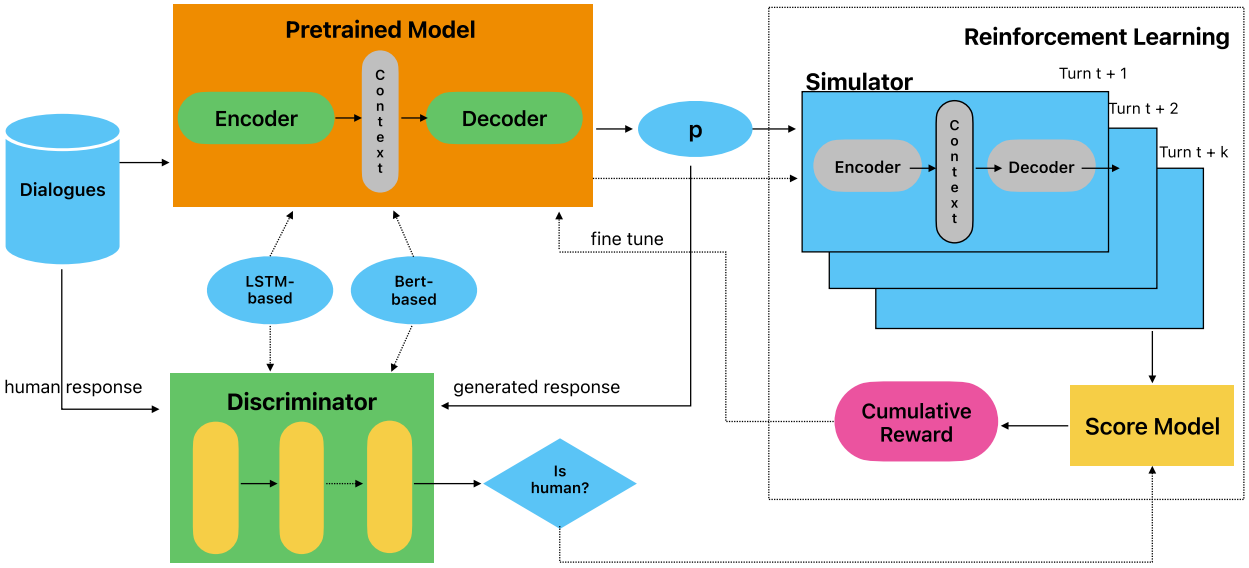

FIGURE 3. Hybrid of Reinforcement and Adversarial Learning for a Chatbot model.

Using the chain rule and relying on the equation $\partial_\theta f(\theta) = f(\theta)\frac{\partial_\theta f(\theta)}{f(\theta)} = f(\theta)\partial_\theta \log f(\theta)$ as in [45], we obtain the following:

$$\begin{aligned}\partial_\theta J(\theta) &= \sum_{a_{t:t+k}\in A} \pi_\theta(a_{t:t+k})\partial_\theta \log \pi_\theta(a_{t:t+k})R(a_{t:t+k}) \\ &= \mathbb{E}_{a_{t:t+k}\sim\pi_\theta}[\partial_\theta \log \pi_\theta(a_{t:t+k})R(a_{t:t+k})] \end{aligned} \quad (14)$$

The algorithm for the proposed model is summarized in Algorithm 3. We use the adversarial learning method to push the model to generate indistinguishable responses from human-generated ones. Using RL as a backbone, we define the quality of the generated responses by their ability to fool the discriminator into believing it is a human response. The output from the discriminator is also feedback to the generator, pushing the model to generate responses to be more like a human.

Algorithm 3 RLGAN-Chat

Require: Input sequence (X), ground-truth output sequence (Y) and history conversation (C_L);

Initialize the Generator network G, and the Discriminator network D
Setup the Response Generator G with pretrained BERT-based
{Generator pre-training with Reinforcement Learning}
Fine-tuning the Generator using reward in (11) on a batch from corpus
{Discriminator pre-training}
while not converged **do**
  (real, fake) ← positive batch from dataset, and negative from G
  Update D using (real, fake)
end while
{Adversarial Training}
for number of training iterations **do**
  for k steps **do**
    (real, fake) ← positive batch from dataset, and negative from G
    Update D using (real, fake)
  end for
  $x \leftarrow$ sample response from training dataset
  $p \leftarrow G(x)$
  $a_{t:t+k} \leftarrow$ simulator(p)
  $r \leftarrow D(a_{t:t+k})$ and (11)
  Calculate the objective function $J(\theta)$ according to (12)
  Update the parameters of network G: $\theta = \theta + \alpha\partial_\theta J(\theta)$
end for

IV. EVALUATION

In this section, we present experimental results and qualitative analysis. Firstly we describe the dataset and explain our choice of the corpus. Then the experimental evaluation of the proposed method is conducted and analyzed against baselines and other works using the same dataset and evaluation metrics.

A. DATA

We use the Daily Dialogue dataset [46] for training and evaluation. The dataset consists of a diverse range of daily life conversations and is divided into three categories: Work (14.49%), Ordinary Life (28.26%), and Relationships (33.33%). It contains over 13,000 multi-turn dialogues covering a variety of daily topics and is collected from various sources for English learners to practice daily conversation.

Recently, many conversation datasets such as the Chinese Weibo dataset [47] and the Twitter Dialog Corpus [48] have

been collected from comments on social networks. Different from real conversations, the language used in social media is often short and noisy. The language in the Dialog Corpus is human-written and more formal. Additionally, the conversations in this corpus typically focus on specific topics and are contextualized. The contents of the conversations are also consistent with real-life experiences because they are often created through collaboration with people. This is why we chose the DailyDialog dataset for our experiments.

B. QUANTITATIVE EVALUATION

Evaluation plays a crucial role in the development of conversational agents. We evaluate dialogue generation systems based on the correlation between human judgment and response generation. The BLEU score [49], a string-matching algorithm, is used as the evaluation criterion. BLEU compares consecutive phrases of the generated response to the consecutive phrases in the reference response, counts the number of matches in a weighted manner, and measures the overlap of words between the generated response and the reference response. A higher score indicates a stronger correlation between human evaluation and response generation. This metric is widely used for evaluating dialogue quality [11], [50]. For the purpose of tokenization and calculating the BLEU score, we utilized the NLTK (Natural Language Toolkit) tool [51] to tokenize the generated responses and compute the BLEU score for evaluation.

We also utilize the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score as the fourth metric, which has been recently adopted for evaluating dialogue quality [52], [53]. This score incorporates various metrics to assess the quality of a response by comparing it to multiple human-generated reference versions. It encompasses n-gram count, word sequence, and word pair to measure the similarity between the output generated by the chatbot and the reference text. To calculate the ROUGE score, we utilized an open-source implementation provided by Google that is widely recognized and used in the research community to evaluate the quality of text-generation tasks.

Recently, BLEURT [54] has been also proposed as a learned evaluation metric and it was used for evaluating text generation tasks' quality [55]. BLEURT is specifically trained to provide continuous scores correlating with human judgments of quality. It considers various aspects of the text, such as fluency, coherence, and semantic accuracy and has been shown to outperform other commonly used metrics, such as BLEU and ROUGE, in capturing the quality of generated text across different tasks. For evaluation using BLEURT score, we employed the latest BLEURT-20 model [56] to obtain accurate and up-to-date assessments of the generated responses according to this metric.

C. EXPERIMENTAL MODELS

All the chatbot models we are building use one core component for encoding input and another for generating text output, and they are typically deep neural networks. There are two modules commonly used for this task, LSTM and BERT, in which LSTM is an early deep-learning architecture known for its efficiency, while BERT is a more recent architecture with outstanding advantages. To evaluate the impact of these architectures on our proposed model in the RL and GAN frameworks, we conducted experiments with both LSTM and BERT.

TABLE 1. Model architecture used in our experiments.

	Reinforcement Learning	Adversarial Learning
LSTM-LSTM		
LSTM-LSTM RL	✓	
LSTM-LSTM GAN		✓
LSTM-LSTM RL + GAN	✓	✓
BERT2BERT		
BERT2BERT RL	✓	
BERT2BERT GAN		✓
BERT2BERT RL + GAN	✓	✓

And then, to demonstrate the effectiveness of the proposed model's features, by comparing experimental results, we built the following models for experimentation, including:

- A baseline model has been created using a Seq2Seq LSTM architecture where the input is the current utterance viewed as a sequence of words and the output is the generated response, which is a sequence of words.
- Another strong baseline which is a BERT2BERT-based model, such as those in [57] and [58].
- We designed to incorporate RL and GAN into the baseline models (namely LSTM2LSTM and BERT2BERT.) To evaluate the influence of each factor, we separately use them, resulting in different experimental models. Notably, for the strategy of solely integrating GAN into the baseline model, we based our design on our previous research [1].
- Our full proposed model combines RL and GAN and uses BERT for both generation and discrimination components.

Table 1 details how we designed the various experimental models. They are built on the following criteria: 1) Use LSTM or BERT for Encoder and Decoder components; 2) Use only RL, or only GAN, or the full model including RL+GAN.

We also summarize the model's architecture of our experiments in Table 1. Moreover, to ensure fairness in comparison, we have set up the parameter configuration similarly to other studies. For all experimental models, the batch size is set to 64 and the learning rate is fixed at 2e-4. We trained the models to minimize cross-entropy using the Adam optimizer [59]. The parameters for the Seq2Seq-based model are based on [60]. We also mapped all out-of-vocabulary (OOV) words to a special token <*UNK*>. The encoder and decoder in the experiments use LSTM structures with 768 hidden neurons and have different sets of parameters. For the models that use pre-trained BERT, we adopt the architecture L=12,

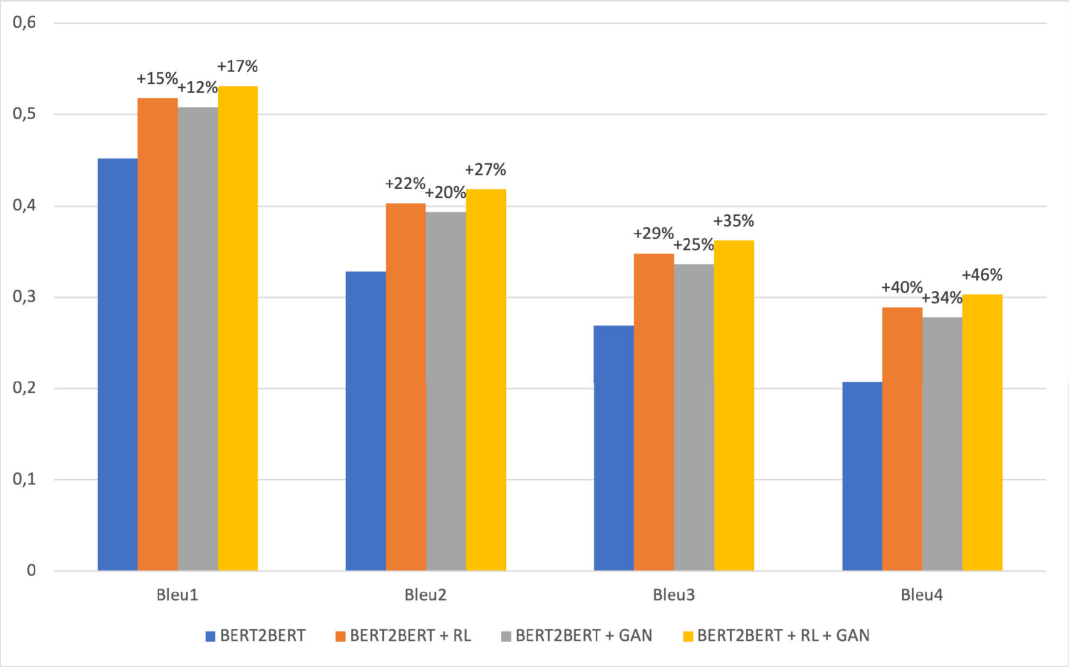

FIGURE 4. Chart showing the performance of proposed models measured by BLEU scores compared with the strong base model (BERT2BERT).

TABLE 2. Results of our experimental models.

Model	BLEU1	BLEU2	BLEU3	BLEU4
LSTM-LSTM	0.185	0.119	0.107	0.086
LSTM-LSTM RL	0.241	0.163	0.151	0.131
LSTM-LSTM GAN	0.218	0.131	0.114	0.089
LSTM-LSTM RL + GAN	0.248	0.164	0.150	0.132
BERT2BERT	0.452	0.328	0.269	0.207
BERT2BERT RL	0.518	0.403	0.348	0.289
BERT2BERT GAN	0.508	0.393	0.336	0.278
BERT2BERT RL + GAN	**0.531**	**0.418**	**0.362**	**0.303**

D=768 for both the Encoder-Decoder model in the Generator and fine-tuning classification in the Discriminator.

D. EXPERIMENTAL RESULT AND DISCUSSION

In this experiment, we used Bleu1, Bleu2, Bleu3, and Bleu4 scores to measure the similarity between the responses generated by the model and the corresponding ground truth utterances in the Test dataset's conversations. Note that we aim for the model to have higher BLEU scores, as this indicates better performance. Table 2 and Fig. 4 display our experimental results. From these results, we have the following observations:

- The use of the BERT component yields significantly better results than using LSTM in all architectures. The obtained BLEU scores are more than twice as high in all equivalent architectures that use BERT instead of LSTM. This can be explained by the fact that BERT's deep network architecture is superior to LSTM. Moreover, when using BERT, we utilized a pre-trained model, which is why BERT performs better by solving the problem of sparse training data.
- The use of RL here is to maximize contextual information of the current utterance when generating responses. When comparing two types of models, single LSTM2LSTM and LSTM2LSTM+RL, we found that adding RL improved BLEU scores by about 30% to 50% (0.241 vs. 0.185 for Bleu1; 0.163 vs. 0.119 for Bleu2; 0.151 vs. 0.107 for Bleu3; and 0.131 vs. 0.086 for Bleu4). The higher index of the BLEU measure, the greater the improvement, indicating that using context through RL has made the generated responses much more accurate.
- When comparing the use of single BERT2BERT and BERT2BERT+RL, we found that adding RL improved BLEU scores by about 15% to 40% (0.518 vs. 0.452 for Bleu1; 0.403 vs. 0.328 for Bleu2; 0.348 vs. 0.269 for Bleu3; and 0.289 vs. 0.207 for Bleu4). It is consistent that the higher index of the BLEU measure, the more effective the use of RL. Furthermore, the fact that a good model (single BERT2BERT) can still be further improved by such a high percentage indicates that the use of RL here is very effective.
- The experimental results when using only GAN to added to LSTM2LSTM and BERT2BERT model without RL are also impressive. When using GAN added to the LSTM2LSTM model, BLEU scores increased by 17.8% for Bleu1 (0.218 vs. 0.185), 10% for Bleu2 (0.131 vs. 0.119), 6.5% for Bleu3 (0.114 vs. 0.107) and 3.3% for Bleu4 (0.089 vs. 0.086). When using GAN for

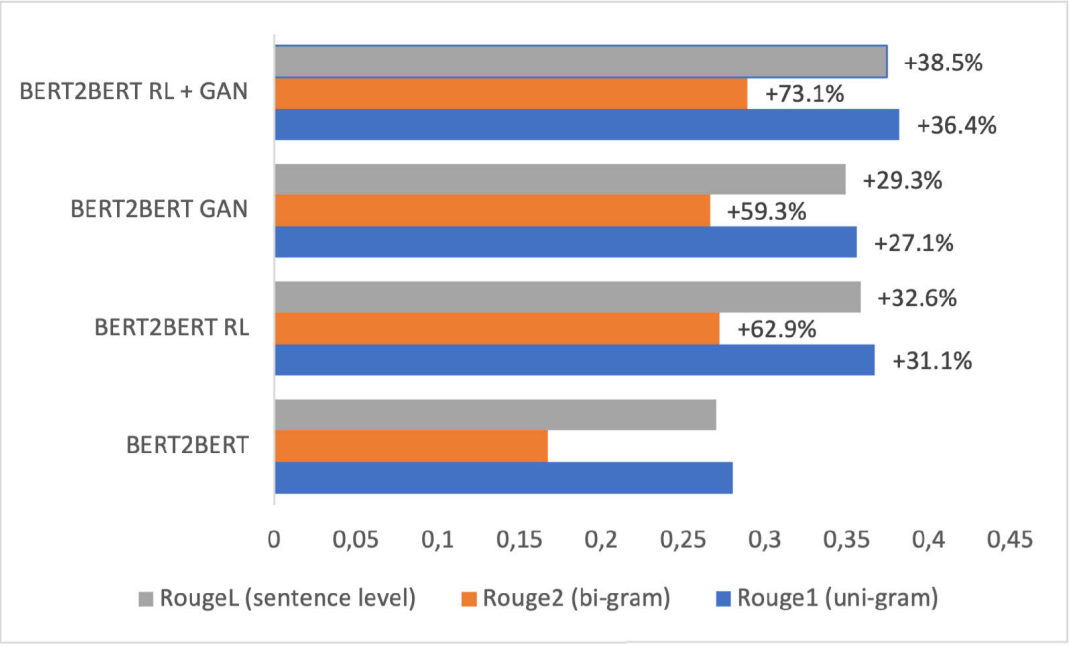

FIGURE 5. Chart showing the performance of proposed models measured by ROUGE scores compared with the strong base model (BERT2BERT).

BERT2BERT model, BLEU scores increased by 12.4% for Bleu1 (0.508 vs. 0.452), 19.8% for Bleu2 (0.393 vs. 0.328), 24.9% for Bleu3 (0.336 vs. 0.269), and 34% for Bleu4 (0.278 vs. 0.207). The special thing here is that as we increase the BLEU index, the improvement of the GAN increases compared to the baseline BERT2BERT, while the improvement of RL decreases. Note that increasing the BLEU score means that the measurement is more based on longer phrases (Bleu4 measures based on length 1, 2, 3, and 4). Therefore, this result further demonstrates that the GAN has improved the naturalness and human-like quality of the generated responses.

- Finally, the full proposed model consists of either the basic LSTM2LSTM or BERT2BERT model combined with RL+GAN. Clearly, the basic BERT2BERT model performs much better than LSTM2LSTM, so when looking at the result table, we should only focus on the BERT2BERT+RL+GAN model. The results achieved show that for all 4 BLEU scores - Bleu1, Bleu2, Bleu3, and Bleu4 - the full model produces the best results. Assuming that the model is an improvement over the BERT2BERT+RL model by adding GAN, we will analyze this improvement. Specifically, the model improved 2.5% for Bleu1 (0.531 vs 0.518), 3.7% for Bleu2 (0.418 vs 0.403), 4% for Bleu3 (0.362 vs 0.348), 4.8% for Bleu4 (0.303 vs 0.289). The improvement is higher for the BLEU score of larger indices, which demonstrates that GAN has improved both the accuracy and naturalness/human-like quality of the BERT2BERT+RL model.

TABLE 3. Results of our experimental models with ROUGE scores.

Model	Rouge1 (uni-gram)	Rouge2 (bi-gram)	RougeL (sentence level)
BERT2BERT	0.280	0.167	0.270
BERT2BERT RL	0.367	0.272	0.358
BERT2BERT GAN	0.356	0.266	0.349
BERT2BERT RL + GAN	**0.382**	**0.289**	**0.374**

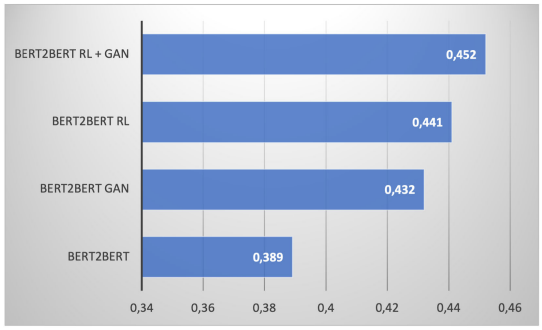

FIGURE 6. Chart showing performance of the proposed models measured by BLEURT scores compared with the strong base model (BERT2BERT).

Fig. 4 presents a visual representation of the BLEU scores results for the 4 experimental models we proposed. Through these results, we can easily see a consistency that the complete model (i.e., BERT+RL+GAN) yields the best results, followed by the BERT+RL model which performs better than the BERT+GAN model, and the BERT model alone yields the lowest results. This is consistent across all 4 BLEU measures (Bleu1, Bleu2, Bleu3, and Bleu4).

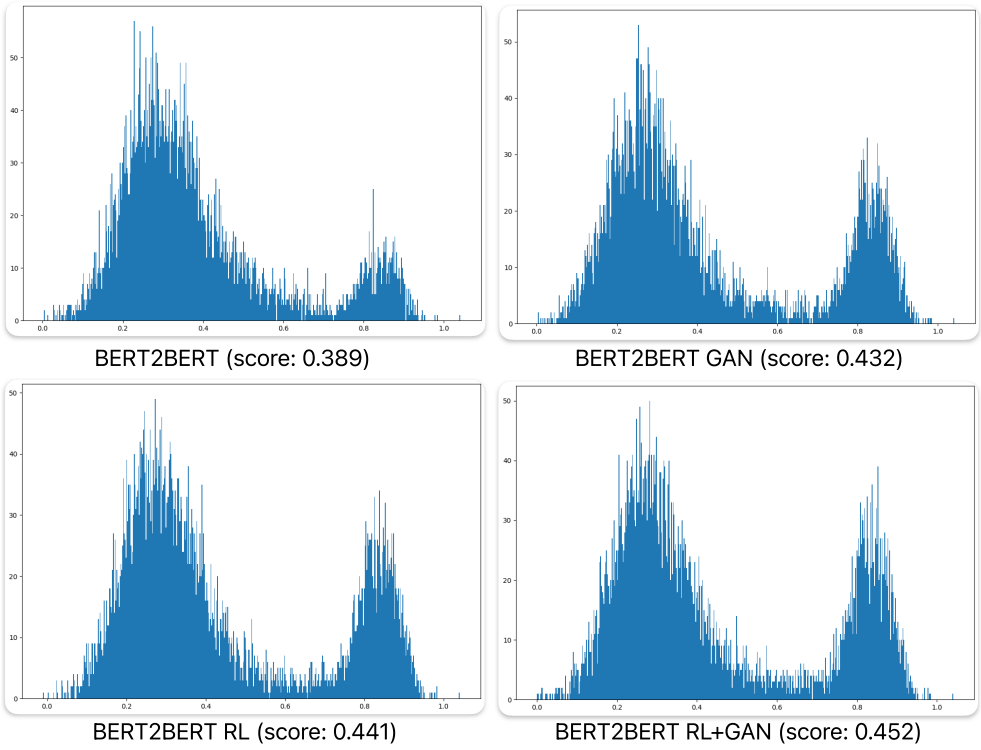

FIGURE 7. Histogram of BLEURT score difference between baseline model (BERT2BERT) and three proposed models.

Our experimental findings have demonstrated superiority of the BERT-based language model over traditional language models. To gain a deeper understanding of the effectiveness of the proposed models, we specifically selected four best models derived from Bert. We compared these models against each other and relevant studies, utilizing evaluation metrics such as ROUGE and BLEURT scores. These comparisons allowed us to comprehensively assess their performance and gauge their alignment with recent research.

As showed in Table 3 and Fig. 5, the four models were compared using ROUGE scores, including uni-gram, bi-gram, and sentence-level metrics. These scores comprehensively evaluate the model's performance in capturing the similarity between the generated outputs and the reference text. When comparing at the word level, the three models, namely BERT2BERT RL, BERT2BERT GAN, and BERT2BERT RL+GAN, demonstrated substantial increases in Rouge1 compared to the baseline model. Particularly, BERT2BERT RL exhibited an increment of 31.1% in Rouge1, while BERT2BERT GAN and the best proposed model (BERT2BERT RL + GAN) achieved 27.1% and 36.4% improvements for uni-gram matches, respectively. Furthermore, when evaluating bi-grams, BERT2BERT RL showed an increase of 62.9%, while BERT2BERT GAN and BERT2BERT RL + GAN exhibited 59.3% and 73.1% improvements in Rouge2 scores, respectively. These findings are strong evidence of the models' enhanced ability to capture uni-gram and bi-gram matches, indicating their effectiveness in generating responses that closely align with the reference text at the word level.

TABLE 4. Comparison of our proposal model with related works.

Model	Bleu1	Bleu2	Bleu3	Bleu4
Proposed model (BERT2BERT RL + GAN)	**0.531**	**0.418**	**0.362**	**0.303**
HRED [46]	0.396	0.174	0.019	0.009
COHA [46]	0.379	0.156	0.018	0.066
COHA + Attention [46]	0.464	0.220	0.017	0.009
Plato [61]	0.486	0.389	-	-

For the sentence level evaluation using RougeL, all three models demonstrated notable improvements compared to the baseline. The RL-based model exhibited an increase of 32.6%, GAN-based model showed a 29.3% improvement, and the best proposed model (RL + GAN) achieved a 38.5% increment in RougeL scores. These advancements underscore the models' effectiveness in generating responses that align well with the reference text at the sentence level. The increases in RougeL scores further validate the improved coherence and similarity of the models' output compared to the baseline.

In summary, the comparisons at the word level indicate that the proposed models surpass the performance of the baseline model, BERT2BERT, by achieving significant increases in Rouge1 for uni-gram matches and Rouge2 for bi-grams. Additionally, the enhancements observed at the sentence

TABLE 5. Sampled conversations generated from the baseline model and the proposed model.

Baseline	Proposed model
Good morning , sir . Is there a bank near here ? Yes, there is one on the corner of this street. Thank you very much. My pleasure. How much is it for a chinese restaurant? It's $ 60.	Good morning , sir . Is there a bank near here ? Yes, there is one on the corner of lincoln street. How long will it take me to get there? Oh, about half an hour. Thank you very much. You're welcome.
What can I do for you today ? I need to return a book. Can i do anything else for you? I would also like to check out this video. Is that all? Yes. Will that be all? Yes. that's it.	What can I do for you today ? I need to return a book. Can i do anything else for you? I want to check this magazine out. We don ' t allow people to check out magazines. Why not? It ' s just a policy of ours. So what am i allowed to check out?
I need to buy a new refrigerator today . Do you want a particular refrigerator? I like a kenmore refrigerator. Would you like a particular refrigerator? I would like a refrigerator. All of them are available. What are they? They are all brand.	I need to buy a new refrigerator today . Were you looking at a particular refrigerator? I like that kenmore refrigerator. This particular refrigerator is a very good choice. Tell me about it. Not only is it affordable, but it comes with all the appliances. What are the appliances? It has an ice maker, water dispenser, and plenty of room on the inside.
Sir , what's wrong with you ? I have a sore throat and headache. Do you have a fever? I don't know. i feel terrible. Let me examine you. don't worry. it's nothing serious. Do you think i should lie in bed? No, not very much. go to see a doctor. Are you seasick?	Sir , what's wrong with you ? I've got a headache and a cough. I see. you've got a cold. What should i do, doctor? Take this medicine three times a day. Yes. Drink a lot of water and have a good rest. you'll be well soon. Thank you.
Whatever would you like to drink ? I'd like a cup of black tea. Would you like some coffee? Yes, i'd like a cup of coffee. Do you want to have coffee or juice? I'll have a cup of coffee. Do you want any orange juice? Yes, please.	Whatever would you like to drink ? I'd like a coke. Forget about the coke. let's drink some wine, shall we? Okay. What would you prefer? red wine or white? Red. What would you like for dinner? I'd like to have a steak sandwich, please.
What are you doing here today ? I'm just shopping. What are you shopping for? I'm shopping for some new clothes. What kind of clothes are you looking for? I'm not sure what kind of clothes to get. You know what you're going to buy? I don't know what to get.	What are you doing here today ? I'm just shopping. What are you shopping for? Some new clothes. how about you? Same here. Have you found anything nice yet? I found a few pairs of pants. Where'd you get those?
Welcome to the McDonald's . What will it be for you , madam ? Two bottles of soya - bean milk and a bottle of red wine. And a bottle of wine? And a bottle of wine? That's right. that's what i wanted to eat. Well... and what did you eat? I went to a movie. Did you eat any food?	Welcome to the McDonald's . What will it be for you , madam ? Two hamburgers, please. For here or to go? To go, please. Tow about fries? A large, with tons of ketchup. And to drink? I'll have a diet coke, please.
How long will it take me to get there ? It's about an hour. I'm afraid i can't go there right now. can you tell me where i can get off? Why not? Where's a bus stop just around the corner. I see. You're right. I think i'm going to get off.	How long will it take me to get there ? It's a short walk. May i take a bus to get there? You can take no. i bus or a taxi. Excuse me, sir. where is the no. 1 bus stop? Oh, go straight on, on the second turning on the right. Thank you very much. So how long does it take?

level, as measured by RougeL, further demonstrate the models' ability to generate responses that closely resemble the reference text in terms of subsequence alignment. These results prove the superior performance of our models in generating coherent and contextually appropriate responses compared to the baseline model.

As shown in Fig. 6, when comparing the BLEURT scores of the four models, BERT2BERT GAN achieved a

11.1% increase in BLEURT score compared to the baseline. BERT2BERT RL exhibited a 13.4% improvement over the baseline. Lastly, BERT2BERT (RL + GAN) demonstrated the most significant enhancement with a 16.2% increase in the BLEURT score compared to the baseline.

To further discuss these improvements, it is important to consider the probability distribution and histogram of the BLEURT scores (as shown in Fig. 7). A histogram provides a visual representation of the distribution of scores, with the x-axis representing the score range and the y-axis representing the frequency or number of occurrences. In this case, the histogram of the BLEURT scores for the four models would indicate the frequency or probability of achieving a particular score. A favorable histogram distribution would show a higher concentration of scores towards 1, indicating a better alignment with the human judgment of quality. A histogram skewed towards higher scores and closer to 1 would indicate a higher likelihood of generating high-quality text.

In summary, the BLEURT scores for the three proposed models demonstrate significant improvements over the baseline model. By examining the percentage improvements and analyzing the probability distribution through a histogram, it can be inferred that these models have achieved better alignment with human judgment and are more likely to generate text with higher quality.

Moreover, we also present a comparison of our proposed model with recent studies on the same experimental dataset (as shown in Table 4). HRED [62] is a hierarchical model based on an encoder-decoder architecture. The model utilizes a sequence of previous queries. COHA [63] builds their model using internal emotional states and captures explicit expressions from predefined emotions. They demonstrate that their proposed model can generate responses that are contextually and emotionally appropriate. PLATO [64] is based on hidden vectors to address the inherent mapping problem in response generation. The model combines bi-directional context and uni-directional characteristics by using attention mechanisms.

The results of our model show superior performance compared to these studies. For previous models in these studies, as the BLEU index increases, for example when comparing Bleu2, Bleu3, and Bleu4 to Bleu1, the value decreases significantly, while our model still produces results that are not too far off. This further confirms that our model generates responses that are accurate (appropriate to the conversation content) and have high natural, human-like tendencies.

We have included some responses generated by the simulator in Table 5, demonstrating that the proposed model can generate meaningful and relevant responses. It is also shown when compared with the strong base model (BERT2BERT model) that the proposed models generate better and more natural conversations. We might think that, through these examples, the integration of full-context modeling into traditional Seq2Seq models, aided by RL, enables the chatbot system to manage information flow in long-term conversations. Our proposed model addresses issues of generating generic responses and maintaining conversational coherence. With its full-context modeling capability and the GAN strategy, the proposed model generates more relevant and natural responses, thereby facilitating sustained conversations.

V. CONCLUSION

In this study, we proposed new models for Chatbot problems that can utilize contextual factors in a conversation. The proposed models combine RL and GAN with a BERT2BERT architecture to capture rich contexts and imitate humans in response generation. The proposed models were designed to generate meaningful and relevant responses in conversations for coherence and naturalness.

We performed a comprehensive set of experiments to assess the proposed model's performance compared to the baselines. These experiments' results showed our approach's effectiveness, as evidenced by significant improvements in the quality of generated responses. The improvements were measured using widely recognized evaluation metrics such as BLEU, ROUGE and BLEURT scores. These results were also consistent with the analysis of the proposed model's characteristics. It is experimentally shown that our developed architectures yielded significantly better results in comparison to recently related studies in the literature. We believe that the proposed model can be incorporated into chatbot systems to improve their quality in practice.

Building on the improved results of this study, our future work will involve the design of additional rewards that align with human decision-making characteristics. By incorporating these rewards into the chatbot's training process, we aim to guide its behavior better to match human expectations, leading to higher-quality interactions and more natural conversations. Additionally, we also plan to explore large language models by incorporating the ideas proposed in this work so as to hopefully enable the development of self-learning conversational agents that can effectively serve various specific purposes, including communication and customer service.

REFERENCES

[1] Q. L. Tran, A.-C. Le, and V.-N. Huynh, "Towards a human-like chatbot using deep adversarial learning," in *Proc. 14th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2022, pp. 1–5.

[2] R. S. Wallace, "The anatomy of A.L.I.C.E." in *Parsing the Turing Test*. Dordrecht, The Netherlands: Springer, 2009, pp. 181–210.

[3] S. Jafarpour and C. J. C. Burges, "Filter, rank, and transfer the knowledge: Learning to chat," Microsoft Research, Redmond, WA, USA, Tech. Rep. MSR-TR-2010-93, 2010. [Online]. Available: https://www.microsoft.com/en-us/research/publication/filter-rank-and-transfer-the-knowledge-learning-to-chat/

[4] Z. Yan, N. Duan, J. Bao, P. Chen, M. Zhou, Z. Li, and J. Zhou, "DocChat: An information retrieval approach for chatbot engines using unstructured documents," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 516–525. [Online]. Available: https://aclanthology.org/P16-1049

[5] R. Al-Rfou, M. Pickett, J. Snaider, Y.-H. Sung, B. Strope, and R. Kurzweil, "Conversational contextual cues: The case of personalization and history for response ranking," 2016, *arXiv:1606.00372*.

[6] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," 2015, *arXiv:1507.04808*.

[7] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Denver, Colorado: Association for Computational Linguistics, May/Jun. 2015, pp. 196–205.

[8] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1192–1202.

[9] T. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2020, pp. 1–25.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[11] O. Vinyals and Q. V. Le, "A neural conversational model," in *Proc. 31st Int. Conf. Mach. Learn.*, 2015, pp. 1–8.

[12] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 654–664.

[13] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, doi: 10.1145/365153.365168.

[14] R. C. Parkison, K. M. Colby, and W. S. Faught, *Conversational Language Comprehension Using Integrated Pattern-Matching and Parsing*. San Francisco, CA, USA: Morgan Kaufmann, 1986, pp. 551–562.

[15] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proc. 25th Int. Conf. Comput. Linguistics (COLING)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 928–939.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, Dec. 2014, pp. 1–9.

[17] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 3506–3510.

[18] S. Kim, O.-W. Kwon, and H. Kim, "Knowledge-grounded chatbot based on dual Wasserstein generative adversarial networks with effective attention mechanisms," *Appl. Sci.*, vol. 10, no. 9, p. 3335, May 2020.

[19] S. Mathur and D. Lopez, "A scaled-down neural conversational model for chatbots," *Concurrency Comput., Pract. Exper.*, vol. 31, no. 10, May 2019, Art. no. e4761.

[20] A. Maeda and S. Moraes, "Chatbot baseado em deep learning: Um estudo para Língua Portuguesa," in *Proc. 5th Symp. Knowl. Discovery, Mining Learn. (KDMiLe)*. Uberlândia, Brazil: SBC, Oct. 2017, pp. 11–18.

[21] A. K. Kushwaha and A. K. Kar, "MarkBot—A language model-driven chatbot for interactive marketing in post-modern world," *Inf. Syst. Frontiers*, pp. 1–18, 2021, doi: 10.1007/s10796-021-10184-y.

[22] S. Sato, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa, "Modeling situations in neural chat bots," in *Proc. ACL, Student Res. Workshop*. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 120–127.

[23] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, Jan. 2022. [Online]. Available: https://www.mdpi.com/2078-2489/13/1/41

[24] A. Kulshreshtha, D. D. F. Adiwardana, D. R. So, G. Nemade, J. Hall, N. Fiedel, Q. V. Le, R. Thoppilan, T. Luong, Y. Lu, and Z. Yang, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.

[25] M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young, "POMDP-based dialogue manager adaptation to extended domains," in *Proc. SIGDIAL Conf.* Metz, France: Association for Computational Linguistics, Aug. 2013, pp. 214–222.

[26] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "POMDP-based statistical spoken dialog systems: A review," *Proc. IEEE*, vol. 101, no. 5, pp. 1160–1179, May 2013.

[27] X. Xiang and S. Foo, "Recent advances in deep reinforcement learning applications for solving partially observable Markov decision processes (POMDP) problems: Part 1—Fundamentals and applications in games, robotics and natural language processing," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 554–581, Jul. 2021.

[28] V. Uc-Cetina, N. Navarro-Guerrero, A. Martín-González, C. Weber, and S. Wermter, "Survey on reinforcement learning for language processing," *Artif. Intell. Rev.*, vol. 56, pp. 1543–1573, Feb. 2023.

[29] Z. Chen, L. Chen, X. Liu, and K. Yu, "Distributed structured actor-critic reinforcement learning for universal dialogue management," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2400–2411, 2020.

[30] S. Verma, J. Fu, S. Yang, and S. Levine, "CHAI: A chatbot AI for task-oriented dialogue with offline reinforcement learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Seattle, WA, USA: Association for Computational Linguistics, Jul. 2022, pp. 4471–4491.

[31] Y.-L. Hsueh and T.-L. Chou, "A task-oriented chatbot based on LSTM and reinforcement learning," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 1, pp. 1–27, Nov. 2022, doi: 10.1145/3529649.

[32] D. Chen, H. Chen, Y. Yang, A. Lin, and Z. Yu, "Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Cedarville, OH, USA: Association for Computational Linguistics, Jun. 2021, pp. 3002–3017.

[33] P. Cai, H. Wan, F. Liu, M. Yu, H. Yu, and S. Joshi, "Learning as conversation: Dialogue systems reinforced for information acquisition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Seattle, WA, USA: Association for Computational Linguistics, Jul. 2022, pp. 4781–4796.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[35] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 3158–3168.

[36] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf., 8th AAAI Symp. Educ. Adv. Artif. Intell.* New Orleans, LA, USA: AAAI Press, 2018, pp. 1–8.

[37] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2157–2169.

[38] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. 31st AAAI Conf. Artif. Intell.* San Francisco, CA, USA: AAAI Press, 2017, pp. 2852–2858.

[39] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.* Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21.

[40] H. Su, X. Shen, P. Hu, W. Li, and Y. Chen, "Dialogue generation with GAN," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–2. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12158

[41] A. Kannan and O. Vinyals, "Adversarial evaluation of dialogue models," 2017, *arXiv:1701.08198*.

[42] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, May 1992, doi: 10.1007/BF00992696.

[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[44] Y. Zhang, Z. Gan, and L. Carin, "Generating text via adversarial training," in *Proc. NIPS Workshop Adversarial Training*, vol. 21, 2016, pp. 21–32.

[45] W. Zaremba and I. Sutskever, "Reinforcement learning neural Turing machines-revised," 2015, *arXiv:1505.00521*.

[46] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint. Conf. Natural Lang. Process. (IJCNLP)*, vol. 1, Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995.

[47] H. Wang, Z. Lu, H. Li, and E. Chen, "A dataset for research on short-text conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Seattle, WA, USA:, USA: Association for Computational Linguistics, Oct. 2013, pp. 935–945.
[48] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Edinburgh, Scotland, U.K.: Association for Computational Linguistics, Jul. 2011, pp. 583–593.
[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.
[50] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 110–119.
[51] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009. [Online]. Available: https://books.google.co.jp/books?id=ScL3wAEACAAJ
[52] J. Kapočiūtė-Dzikienė, "A domain-specific generative chatbot trained from little data," *Appl. Sci.*, vol. 10, no. 7, p. 2221, Mar. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/7/2221
[53] R. Yang, Z. Li, H. Tang, and K. Zhu, "ChatMatch: Evaluating chatbots by autonomous chat tournaments," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7579–7590. [Online]. Available: https://aclanthology.org/2022.acl-long.522
[54] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Cedarville, OH, USA: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704
[55] S. H. Kumar, H. Su, R. Manuvinakurike, M. C. Pinaroc, S. Prasad, S. Sahay, and L. Nachman, "Cue-bot: A conversational agent for assistive technology," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 196–203. [Online]. Available: https://aclanthology.org/2022.acl-demo.19
[56] A. Pu, H. W. Chung, A. P. Parikh, S. Gehrmann, and T. Sellam, "Learning compact metrics for MT," in *Proc. EMNLP*, 2021, pp. 1–13.
[57] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 264–280, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.18
[58] C. Chen, Y. Yin, L. Shang, X. Jiang, Y. Qin, F. Wang, Z. Wang, X. Chen, Z. Liu, and Q. Liu, "bert2BERT: Towards reusable pretrained language models," 2021, *arXiv:2110.07143*.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.
[60] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1–10.
[61] J. Fan, L. Yuan, H. Song, H. Tang, and R. Yang, "NLP final project: A dialogue system," Hong Kong Univ. Sci. Technol. (HKUST), 2020.
[62] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder–decoder for generative context-aware query suggestion," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.* New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 553–562.
[63] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf., 8th AAAI Symp. Educ. Adv. Artif. Intell.* New Orleans, LA, USA: AAAI Press, 2018, pp. 730–738.
[64] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Cedarville, OH, USA: Association for Computational Linguistics, Jul. 2020, pp. 85–96.

QUOC-DAI LUONG TRAN received the B.S. and M.S. degrees in information technology from National University Ho Chi Minh, Vietnam, in 2009 and 2013, respectively. He is currently pursuing the Ph.D. degree in computer science with Ton Duc Thang University (TDTU), Ho Chi Minh City, Vietnam. Since 2018, he has been a Lecturer with the Department of Computer Science and a member of the Natural Language Processing and Knowledge Discovery Laboratory, TDTU. His current research interests include natural language processing, deep learning, reinforcement learning, and adversarial learning.

ANH-CUONG LE received the bachelor's and master's degrees in information technology from Vietnam National University, Hanoi, in 1998 and 2001, respectively, and the Ph.D. degree in information science from the Japan Advanced Institute of Science and Technology (JASIT), in 2007. He is currently an Associate Professor of computer sciences and the Head of the Natural Language Processing and Knowledge Discovery Laboratory, Ton Duc Thang University (TDTU), Ho Chi Minh City, Vietnam. He is also responsible for several industrial projects of opinion analysis, text summarization, and chatbot. His current research interests include natural language processing and machine learning.

VAN-NAM HUYNH (Member, IEEE) received the Ph.D. degree in mathematics from the Vietnam Academy of Science and Technology, in 1999. He was a Visiting Professor with the National Electronics and Computer Technology Center, Thailand, in 2019; an Adjunct Professor with Chiang Mai University, Thailand, from 2015 to 2017; and a part-time Lecturer with Tsukuba University, Japan, from 2011 to 2015. He is currently a Professor with the School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST). His current research interests include decision analysis and management science, AI and machine learning, modeling and reasoning with uncertain knowledge, argumentation, multi-agent systems, and Kansei information processing and applications. He currently serves as an Area Editor for *International Journal of Approximate Reasoning* and *Array* and the Editor-in-Chief for *International Journal of Knowledge and Systems Science*.