Imperial College
London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Target-Guiding for Open-Ended Chatbot Interactions in Digital Psychotherapy

Author:
Philip Nag

Supervisor:
Dr. Anandha Gopalan

Submitted in partial fulfillment of the requirements for the MSc degree in Computing Science of Imperial College London

31. August 2022

Abstract

The current mental health crisis has underscored the need for novel forms of delivering psychotherapy, such as chatbots. To provide a therapy known as Self-Attachment Technique (SAT), the Algorithmic Human Development Group at Imperial College London developed the SAT-Chatbot. When patients first interact with the SAT-Chatbot, they engage in chit-chat, during which the bot has the objective of capturing the patient's emotional state. This novel research applies Reinforcement Learning with Proximal Policy Optimization to teach the SAT-Chatbot a conversation strategy that guides the dialogue to this objective. This training is highly effective, enabling the SAT-Chatbot to capture patients' emotions efficiently in up to 86% of conversations. As a positive side effect, it also significantly improves the smoothness, engagement, and empathy of the dialogue compared to previous SAT-Chatbot implementations.

Acknowledgments

With no previous experience in Natural Language Processing or Reinforcement Learning before the start of this project, completing this research was highly challenging and would not have been possible without the support of key members of the Imperial College Community.

I would like to express my deepest gratitude to:

- **Dr. Anandha Gopalan** - for your close guidance throughout the entire project, your challenging questions, pragmatism, and for always being available in case I had additional inquiries
- **Dr. Marek Rei** - for your immediate and very helpful feedback in the early stages of the research, and your proactivity in providing additional advice for the Individual Project
- **Prof. Abbas Edalat** - for lending me an ear and providing stimulating ideas during the assessment of the chatbots and for further work
- **Neophytos Polydorou, Lisa Alazraki, Hongyuan Yan** - for your support in introducing the SAT-Chatbot to me and helping me shape my approach

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Depression has been recognized as the leading cause of disability globally [1] in what many public health officials have declared a *mental health crisis*, amplified by factors such as COVID-19, the war in Ukraine, and social media [2]. There is tremendous pressure on the public health system, with recent studies noting that at least 1.5 million people in England were awaiting mental health treatment in 2021 [3]. It has resulted in calls for novel forms of providing psychotherapy, such as digital psychotherapy [4].

Self-Attachment Technique (SAT) is a novel form of psychotherapy developed to foster positive emotions by creating a strong bond between a mental health patient and their childhood self. Such Therapy is self-administered and relies on patients engaging in behavioral protocols that depend on their current emotional state [5]. Given the promising advances in digital psychotherapy and AI-enabled Language Models (LM), the Algorithmic Human Development (AHD) Group at Imperial College London has developed the SAT-Chatbot. It is designed to engage naturally with patients, capture their emotional state, and recommend an appropriate SAT protocol to them [6]. A development area in the latest iteration of the SAT-Chatbot is the fact that parts of the chat interaction, where users engage in chit-chat with the bot, do not follow an explicitly defined dialogue strategy [7], despite having the aim of capturing the user's emotional state.

The objective of this research is to introduce target-guiding behavior to the SAT-Chatbot. Specifically, the chatbot's LM is taught a dialogue strategy to effectively achieve the target of the conversation, capturing the emotional state of the patient, whilst maintaining the smoothness of the interaction. This objective is implemented by using Proximal Policy Optimization (PPO), an algorithm in Reinforcement Learning (RL), to train the SAT-Chatbot's Language Model.

This paper first reviews the current state of research on digital psychotherapy, Language Models, Reinforcement Learning, and target-guiding for dialogue systems. Together with insights from preliminary experiments applying PPO to conversational agents, the Reinforcement Learning pipeline is defined. Three alternative models for achieving target-guiding are defined, trained, and evaluated using an automized

assessment and a non-clinical human trial.

The final optimized model is highly effective in reaching the target of the conversation, achieving the objective efficiently in up to 86% of conversations. It not only maintains but strongly improves the conversational smoothness and level of empathy compared to previous SAT-Chatbot implementations.

Chapter 2

Background

This section first reviews the current state of Digital Psychotherapy, including applications of Natural Language Processing (NLP), before introducing SAT and the design of the SAT-Chatbot. To better inform the final research method, it continues by putting the SAT-Chatbot into the context of common dialogue system approaches and outlines the potential of RL and PPO.

2.1 Digital Psychotherapy

Digital psychotherapy is heralded by many as having the potential of disrupting mental health care delivery [4] and is a key component of several government-level healthcare system strategies, such as the *NHS Long Term Plan* [8]. It entails supporting the delivery of psychotherapy to a patient using e.g. video-/audio-based interactions, automized chatbots, or immersive virtual reality environments [9]. This promises to increase the ease of access to psychotherapy for patients, especially in regions with low therapist availability, since it can be efficiently scaled to new users at negligible incremental costs [4].

NLP has been applied in myriad use cases in digital psychotherapy, most predominantly in the analysis of therapy transcripts to improve mental health care outcomes [e.g. 10]. First psychotherapeutic uses of chatbots, which are designed to simulate a counterpart's responses in a chat conversation, date back to as far as 1966 [11]. More recent advances in NLP have fuelled the development of more sophisticated chatbots: In a recent survey, Xu and Zhuang note that such chatbots fulfill a variety of different purposes when chatting with the patient, including providing advice, displaying compassion, making the patient laugh or engaging the patient in exercises [12]. Whereas current chatbot implementations in digital psychotherapy such as *Evebot* [13] and *TeenChat* [14] have been shown to perform well, there remains further potential by widening the set of use cases for such chatbots and by improving the quality of their response generation [12].

2.2 Self-Attachment Technique and the SAT-Chatbot

Self-Attachment Technique is a novel form of psychotherapy that aims to improve a patient's mental health by helping them form a strong sense of attachment between their adult and child selves [5]. It is based on the Attachment Theory in developmental psychology, which has identified a patient's insecure bond with their parents during childhood as a significant cause of mental health problems later in their life [15]. Two distinguishing characteristics of SAT open it up to novel forms of delivery, including digital psychotherapy: the fact that it is self-administered, and that it relies on well-defined protocols [5]. There are 20 such protocols, which are exercises that the patients perform by themselves. The choice of exercise depends on the patient's current emotional state, e.g. whether the person is 'happy/content' or 'anxious' [6]. An overview of these protocols can be found in Appendix A.

SAT-Chatbot

The SAT-Chatbot was designed by the Algorithmic Human Development Group at Imperial College London with the objective of interacting empathetically with the patient and recommending the appropriate SAT protocols to them [6]. In the latest iteration of the chatbot by Yan [7], which builds on previous work by Alazraki et al. [6], the conversation proceeds in two stages:

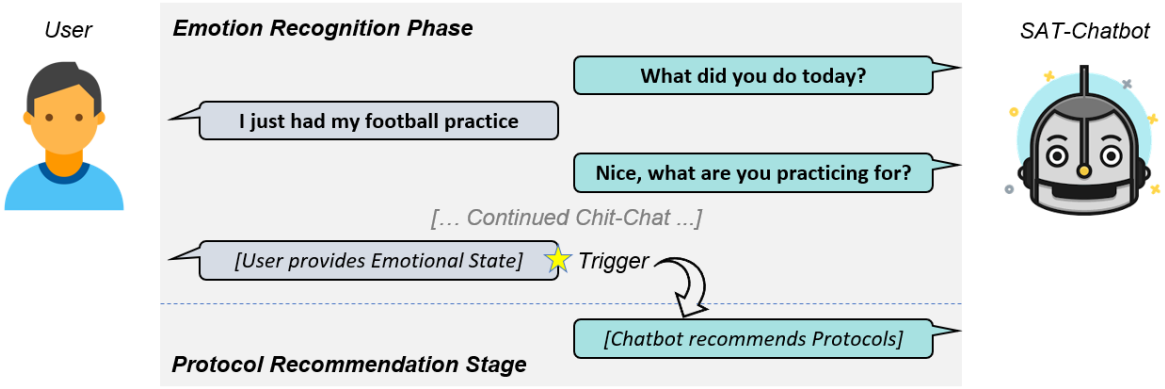


Figure 2.1: Two Phases in SAT-Chatbot Interaction

1. **Emotion Recognition Stage**: The user and the chatbot engage in chit-chat, throughout which the chatbot displays empathy to the user. The objective of this stage is to capture the general emotional state of the user. Once this is achieved, the conversation proceeds to the Protocol Recommendation Stage.
2. **Protocol Recommendation Stage**: This stage has the objective of suggesting concrete protocols to the user. If the general emotion captured in the Emotion Recognition Stage was negative, the chatbot asks additional clarifying questions to narrow down the set of relevant protocols. The bot then displays recommended protocols for the user to choose from.

A key pain point in the current iteration of SAT-Chatbot's Language Model lies in the fact that the Emotion Recognition Stage lacks a profound dialogue strategy to ensure

that the user reveals their emotional state. To prevent the Emotion Recognition Stage from continuing with generic chit-chat forever and never reaching the Protocol Recommendation Stage, the SAT-Chatbot currently forces asking the patient about their emotions after 7 turns[1] [7]. This rule does not take into account how well this question fits into the current conversational context.

The technical implementation of the SAT-Chatbot and its performance is outlined in greater detail in Section 3.1.

2.3 Language Models and Dialogue Systems

Language Models, a key concept in Natural Language Processing, statistically determine the probability of observing a sequence of words [16]. When generating text, these models are used to predict the next word in a sequence, by taking all previous words of the sequence as input and determining the conditional probabilities of observing candidate continuing words from a set vocabulary [17]. Most current LM implementations are neural and early breakthroughs used Recurrent Neural Networks, which were more effective for language generation than previous non-neural n-gram-based methods [16].

Most state-of-the-art LMs implement a Transformer neural network architecture, which, during encoding, pays selective attention to specific sections of the input before feeding it through the network [18]. Radford et al. showed that such architectures can be trained effectively for a variety of tasks through unsupervised pre-training on large text corpora, before fine-tuning the model for specific NLP applications [19]. Popular and publicly available examples of such pre-trained Transformer models include GPT-2, developed by OpenAI [19], and BERT, developed by Google [20].

Dialogue Systems Design

Language Models are employed for a number of NLP tasks, including translation, classification, and dialogue systems [21]. Dialogue systems, such as the SAT-Chatbot, are agents that converse with a user in written language over a number of turns. Chatbots are classified either as *task-oriented* or *open-domain* [16]: *Task-oriented* agents interact with the user with the sole purpose of achieving a conversational objective, e.g. making a restaurant booking or answering a question on a website. *Open-domain* chatbots are designed to converse naturally with the user in a chit-chat manner, without following a conversational objective [22]. Hybrids of these conversational agents exist [e.g. 23] and, as a key focus of this research, will be explained in greater detail in Section 2.5.

The type of dialogue system often informs the LM's technical architecture to produce an utterance. Jurafsky et al. explain that chatbots can use *template-based retrieval*, where the agent selects an utterance from a corpus of candidate utterances. This

[1] A turn is defined as an utterance produced either by the user or the chatbot.

enables a higher degree of control over the possible content of utterances and is therefore often employed in *task-oriented* agents [16]. Alternatively, the chatbot can be designed to produce an utterance through the LM's *text generation*, i.e. producing a response to the user's input token-by-token based on the previous conversation history. This method produces more flexible and relevant responses to a wider range of user utterances and is thereby the primary method for *open-domain* agents [16; 24].

Language Model Assessment

Having an accurate metric to measure language quality is important both for the training and evaluation of Language Models. One of the most common metrics is *perplexity*, which is defined as the inverse of the conditional probabilities of observing a generated sequence of words, and the objective is commonly to minimize this score [16]. Other popular metrics such as BLEU [25] and ROUGE [26] that are specific to certain NLP tasks, e.g. translation or summarizing, exist and often require golden-truth text to be computed. Importantly, a growing body of research points to a lack of correlation of existing language quality scores with human judgment [e.g. 27; 28]. As a result, research involving language generation should also capture human feedback to evaluate language quality [29].

For dialogue systems specifically, measuring the semantic similarity, i.e. the topic-relatedness of text, provides a useful metric for the smoothness of a conversation. Current approaches, which often measure the cosine distance of word embeddings [30] or use BERT classifiers to gauge similarity [31], have been successfully applied in the development of conversational agents [32; 33]. A leading implementation is Cer et al.'s *Universal Sentence Encoder* which encodes words, sentences, or paragraphs into 512-dimensional vectors and has demonstrated very strong performance on myriad NLP tasks [34]. Using this encoding to compute semantic distance in conversations presents a promising prospect for the fine-tuning of conversational agents. As a final point on dialogue assessment, generating a sufficient quantity and breadth of interactions to assess a dialogue system is highly labor-intensive, which is why simulators that emulate user responses are often used to interact with the dialogue system for evaluation. The choice of such user simulators should be specific to the domain of the dialogue system in order to best mimic user behavior [35]. This approach has been a successful component of many dialogue system platforms, such as ConvLab [36].

2.4 Reinforcement Learning for Language Models

Reinforcement Learning follows the objective of teaching an agent a desired behavior in an environment, which it learns by optimizing its return of a reward signal [37]. The subject of the optimization is the agent's policy π, which takes its perceived state of the environment x as input and dictates the agent's action a. At each training step, the agent takes x and approximates the expected reward r that actions a contingent

on its policy π would produce. The expected long-term return of a policy is known as the *value function*. When an action a_t is performed and the actual reward signal r_t is received, the algorithm uses this input in order to refine either the policy directly (known as *policy-based optimization*) or the estimate of the value function (known as *value-based optimization*) [37]. Deep Reinforcement Learning refers to the use of neural networks for these approximations [22]. Compared to Supervised Learning, the major advantage of RL is that it is not reliant on annotated data sets for training, which are often difficult to source. In addition, the performance of Supervised Learning models by definition can at most match the quality of the annotated data sets they are trained on, whereas RL models with well-defined fitness functions may further outperform. However, the reward scheme for RL training must be well defined to achieve the desired behavior [37].

Reinforcement Learning has been successfully applied in many domains such as robotics and health care [21]. RL advances at the intersection with Natural Language Processing were initially slow [38] but gained traction especially for the training of policy for dialogue systems [21]. Li et al. note that end-to-end training approaches for dialogue systems often sought to strongly limit the agent's action space, e.g. by employing template-based retrieval strategies for utterance generation [21]. This was in an effort to improve the stability of the optimization, which suffers from large output spaces such as all the words in a vocabulary chosen across an utterance of unrestricted length [38]. Such implementations enabled applications of RL to train dialogue systems for *task orientation*, which is further elaborated upon in Section 2.5.

With the advent of large pre-trained generative Language Models, the main focus in research turned to the fine-tuning of dialogue policy most commonly using supervised and unsupervised learning [19], instead of RL. Any Reinforcement Learning algorithm for dialogue policy with generative models must ensure that its pre-trained low perplexity losses are maintained whilst the agent exploits behaviors that maximize its reward [38]. This challenge is compounded by the aforementioned fact that current metrics for language quality of generative models, such as perplexity, do not correlate well with human assessment [28], which questions their use for training with Reinforcement Learning.

As a final note, Chen et al. recently underscored Transformer models' potential as a network architecture for general Deep Reinforcement Learning, also outside of the NLP domain. This stems from the scalability of the Transformer infrastructure, which allows it to make predictions over long time horizons and enables it to perform well on sequence-modeling tasks with long-range dependencies such as in NLP. They demonstrate that this scalability likewise lends itself well to policy gradient estimations and long-term reward predictions [39].

Proximal Policy Optimization and its Applications in NLP

Proximal Policy Optimization is a Reinforcement Learning algorithm introduced by Schuman et al. that is targeted, among others, at increasing the stability of conven-

tional policy-based optimization methods. When performing updates on the policy π, its objective is to optimize the behavior of the agent whilst ensuring that the deviation from a reference policy ρ is small. It does so by clipping updates to policy π to a small range [40]. Ziegler et al. introduce a further control mechanism to PPO by implementing an additional penalty term to the reward function which reflects the KL-Divergence, i.e. the Kullback–Leibler-Divergence[2] between the Log probabilities of π and ρ. They weigh this term by KL-Coefficient β to produce the following reward function R, where the user-defined reward before applying the penalty term is r [41]:

$$R(a, x) = r(x, a) - \beta \times log \frac{\pi(a|x)}{\rho(a|x)}$$

Ziegler et al. were also among the first to suggest the use of Proximal Policy Optimization for the fine-tuning of text generation. The approach was trialed by capturing sparse human feedback as the reward signal and was shown to perform well on text continuation and summarizing tasks with GPT-2 Transformer models. Importantly, the user-defined reward signal r only rewarded the model for fulfilling the NLP task, and the authors credited the KL-penalty for maintaining coherence and language quality [41].

A year later, Faal et al. demonstrated the potential for applying Proximal Policy Optimization to train Transformer-based dialogue systems. Their research was targeted at improving the level of engagement of the Language Model by rewarding the agent for responses with high informational relatedness to user prompts, an approach that performed strongly compared to baselines [42].

2.5 Target-Guided Dialogue

As explained in Section 2.2, in the case of the Emotion Recognition Stage of the SAT-Chatbot, a specific objective is imposed on an open-domain conversation between the patient and the agent. In the taxonomy introduced in Section 2.3, this requires a hybrid implementation of a *task-oriented* and *open-domain* conversational agent that guides a chit-chat interaction to a pre-defined target. Tang et al. define that such hybrid agents must address two concurrent objectives in their dialogue policy [32]: First, they must ensure the objective of the conversation, in this case identifying the patient's emotional state, is achieved. Second, they must continue to interact with the patient in a smooth and engaging manner, i.e. choosing dialogue topics that are relevant to the conversation history [32]. The latter requirement calls for using target-guided response *generation*, as opposed to pure *template-retrieval* (see Section 2.3). All in all, target-guiding behavior for open-domain systems with conversational goals is still a niche body of research that lacks large-scale data sets which would enable supervised training, as underlined by Gupta et al. [27].

Currently, combinations of rules- and neural network-based approaches for target-

[2]KL-Divergence is a standardized measure of the difference between two probability distributions given the same input x [41].

guiding are dominant and have perform well for achieving simple conversational targets, such as uttering keywords or sentences, whilst maintaining the smoothness of a chit-chat conversation [e.g. 27; 32; 33]. The majority of these approaches progressively move the conversation closer to the conversational objective, in what can be characterized as a *multi-turn strategy* [32; 33]. Tang et al. address the objectives of task success and conversational smoothness explicitly in their research. Their implementation takes extracted keywords from the conversation history as input and pre-selects candidate keywords that have a smaller semantic distance to a target keyword compared to the current topic of conversation. They measure semantic distance using the cosine similarity of word embeddings. From these candidate keywords, a separate neural network selects the final next utterance topic, i.e. the keyword with the highest conversational smoothness, which is then used by the Language Model to generate a response utterance [32]. Qin et al. further improve smoothness by explicitly considering knowledge relations between the keywords in the conversation history [33], which also found application in other target-guiding strategies such as the use of symbolic knowledge graphs by Wu et al. [43].

Single-turn target-guiding strategies also exist, and a very recent approach by Gupta et al. offers an interesting structure for approaching the problem [27]. In their research, where the objective is defined as uttering a target sentence, they optimize for generating the smoothest *bridging sentence* that leads to the target over a single turn [27]. The final response to a user prompt is a two-sentence utterance comprised of this *bridging sentence* and the target sentence. Optimizing the generation of such *bridging sentences* could present a simple yet effective approach for target-guiding with Reinforcement Learning, which is further elaborated upon in Section 3.3.2.

Reinforcement Learning has been a popular approach to training pure *task-oriented* systems, which do not have conversational smoothness requirements to satisfy. Hence, none of these approaches reward the agent for choosing conversation topics with close semantic similarity [44; 45; 46; 47; 48]. Example applications of this include making movie bookings [45; 46; 48] or retrieving information from a database [49]. The appeal of RL in this context stems from its ability to train dialogue policy to reach more complex targets. Peng et al. demonstrate this by training an agent to obtain information from the user to achieve a combination of inter-dependent subtasks concurrently, e.g. reserving a hotel and booking a flight whilst allowing enough time for commuting [45]. Another example is provided by Zhou et al. [44], who demonstrate RL's ability to achieve complex multi-targets, such as checking a user's eligibility for a refund. As a final point, most RL implementations for task-oriented systems were not applied to large-scale pre-trained generative Language Models [44; 45; 46; 47; 48] introduced in Section 2.3.

Chapter 3

Research Objective and Method

The objective of this research is to introduce target-guiding to the Emotion Recognition Stage of the SAT-Chatbot. Specifically, this entails fine-tuning its current Language Model to effectively reach the target of the conversation, i.e. having the user reveal their emotional state, whilst providing a high conversational smoothness.

Having reviewed previous research for target-guiding in dialogue systems, as well as the application of different Machine Learning approaches in NLP, Reinforcement Learning with Proximal Policy Optimization presents an interesting prospect to train dialogue policy of a hybrid *open-domain* dialogue system with conversational goals. The allure of RL is driven in part by the lack of large-scale data sets for target-guiding [27] that would have enabled fine-tuning with unsupervised or supervised learning, and RL's potential to train for more complex conversational objectives [e.g. 44; 45] than the keyword or sentence-based goals of current target-guiding implementations [27; 32; 33]. Given the requirement of such dialogue systems to maintain high conversational smoothness, the ability of PPO to train for objectives while preventing large divergences from an original policy appears especially attractive.

Hence, in the following, the application of Proximal Policy Optimization is analyzed to train the dialogue policy of the SAT-Chatbot to reach its conversational objective effectively and smoothly. Three different approaches for achieving task success and conversational smoothness are compared, as will be explained in further detail in Section 3.4. The performance of the trained conversational agents is evaluated through automized assessments of chat simulations and a human trial, as outlined in Section 3.5. In explicit terms, this project contributes novelty to the existing body of research by applying Reinforcement Learning with Proximal Policy Optimization to fine-tune an *open-domain* dialogue system with conversational goals for target-guiding. The code is made publicly available under `https://gitlab.doc.ic.ac.uk/pn21/target-guiding-sat-chatbot`.

3.1 Baseline Model

The Language Model employed for the Emotion Recognition Stage in the most recent iteration of the SAT-Chatbot, developed by Yan [7], is used as the baseline for this research. It will serve as the model that is further fine-tuned to introduce target-guiding behavior and against which the final performance of the trained models is evaluated.

The current LM architecture is a pre-trained GPT-2 Transformer model which has been fine-tuned for specific behaviors in the context of the SAT-Chatbot's Emotion Recognition Stage. At its base sits Open AI's GPT-2 model with 1.5 billion parameters, pre-trained on 40GB of text data [50]. It is sourced using HuggingFace's *Transformers* library and encodes inputs into vectors of a maximum of 1024 tokens, using a base vocabulary of 50,257 tokens [50].

It has been fine-tuned to elicit the following behavior:

- **Engaging in Dialogue**: The model was fine-tuned in an unsupervised manner on a large chit-chat dialogue corpus, with the objective of learning dialogue structure and how to respond to another party in an engaging manner. To support this, additional tokens were added to the vocabulary to increase its overall size to 50,261 [7].
- **Displaying Empathy**: The model was further fine-tuned in an unsupervised manner on a data set of dialogues centered around emotional scenarios. This step was designed to teach the model to display emotional concern and respond empathetically [7].

Whereas it produces solid results on relatedness to user prompts for *single*-turn responses, with survey respondents ranking them at 4.4/5.0 [7], the baseline model's performance over *multi*-turn conversations varies strongly and relatedness is often poor. In a non-clinical trial where participants engaged in conversation with the chatbot, 70% of respondents could *not* agree with the statement 'When I interact with the chatbot, I found the conversation to be engaging'[1] [7]. The same gaps in relatedness over multi-turn interactions are also observed when experimenting with the baseline model.

These low levels of response-relatedness in the baseline should be controlled for when evaluating the performance of the models developed in this research.

3.2 Data

In order to introduce target-guiding behavior to the SAT-Chatbot's Language Model, a dialogue data set is required to both form an environment in which to optimize the model using Reinforcement Learning and to assess the developed models thereafter.

[1] Out of 10 respondents, 4 people *Strongly disagreed* or *Disagreed* with this statement. 3 people *Neither disagreed nor agreed* and only 3 people *Agreed* or *Strongly agreed* [7].

Since the target Language Model is, among other aspects, developed to engage users in chit-chat conversations, the data set should be comprised of authentic general-purpose human chat transcripts that cover a wide variety of topics.

DailyDialogues, developed by Li et al., is a data set of 13,118 multi-turn English dialogues [51] and serves this purpose well. It is comprised of *hand-written* English utterances and is, hence, more representative of human-to-human conversations and less susceptible to noise [51] than *crawled* data sets gathered from websites such as Twitter [52] or Weibo [53]. The data is split into training (80%, 10,494 conversations), validation (10%, 1,312) and test sets (10%, 1,312). The use of this data for model training is explained in greater detail in Section 3.3, whilst its use for assessment is outlined in Section 3.5.

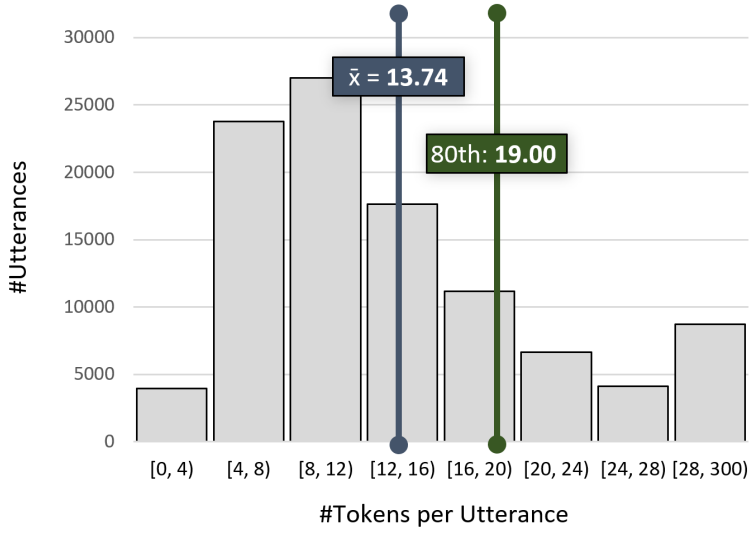


Figure 3.1: Utterance Length in *DailyDialogues*

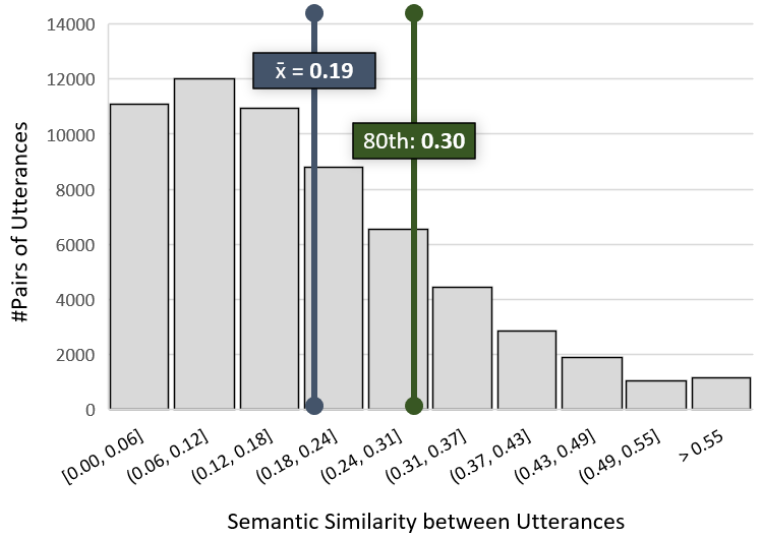


Figure 3.2: Semantic Similarity between Response Utterances in *DailyDialogues*

Additionally, the *DailyDialogues* data set also enables distilling important insights

about natural human conversation that inform the setup of the Language Model development in this research. Specifically, Figure 3.1 shows that an average utterance length is 13.74 tokens, with 80% of utterances consisting of 19.00 tokens or less. Figure 3.2 demonstrates that the average semantic similarity of sentences within the same utterance, as measured by the inner product of sentence encodings using the *Universal Sentence Encoder* [34], is 0.19, with 0.30 representing the 80th percentile.

3.3 Reinforcement Learning Pipeline

This section develops the PPO-based pipeline for training the SAT-Chatbot's Language Model. It reviews insights drawn from early experiments with Proximal Policy Optimization and open-ended dialogue systems to define the core RL components of this approach, including action, state, policy, rewards and optimization logic. These components are the foundation on which three alternative Language Model development approaches are later formulated in Section 3.4.

3.3.1 Experiments with Proximal Policy Optimization

As mentioned in Section 2.4, the existing body of research on applying Proximal Policy Optimization to dialogue systems is scarce, and there exists no previous application of PPO for target-guiding in dialogue systems. Hence, a wide-ranging series of experiments is performed to explore promising RL-PPO configurations for this research, including pipeline structure, reward definition and hyperparameter settings. The types and ranges of the experiments are inspired by the two most closely related pieces of research to this work, contributed by Ziegler et al. [41] and Faal et al. [42]. The insights from these experiments are shared in part to motivate some of the design decisions of the final Reinforcement Learning setup, and also to provide guidance to others who wish to conduct further research on the topic of PPO for conversational agents.

Taking from Sections 3.1 and 3.2, all configurations will be targeted at training the *agent*, the baseline GPT-2 Language Model currently in use during the Emotion Recognition Stage of the SAT-Chatbot. The agent will interact with its *environment* which is a user prompt taken from a large-scale human dialogue data set described in Section 3.2, and produces an utterance that is scored using a defined *reward scheme*.

The following was analyzed over 340 training runs with distinct PPO configurations:

1. **ACHIEVING TASK SUCCESS**: Target representations in the agent's and the user's actions; Reward metrics to signal target proximity
2. **ACHIEVING CONVERSATIONAL SMOOTHNESS**: Rewards to increase relatedness of dialogue and quality of English; Changes to the Language Model
3. **REWARD WEIGHTING**: Sensitivities of reward scaling and weights
4. **HYPERPARAMETER SETTINGS**: Learning rate, KL-Coefficient, training steps

Throughout these experiments, it was closely investigated if and after how many steps the configuration causes the model to converge on a desired behavior.

Observations during Experimentation

The following observations are drawn from these experiments:

1. **ACHIEVING TASK SUCCESS: Optimization most effective with Target Representation in Agent's Actions, using Semantic Similarity**

   Since the conversational objective of the Emotion Recognition Stage, capturing the patient's emotional state, is uttered by the user and not the agent, first experiments included configuring a user simulator to generate responses to bot utterances. For every utterance the bot produced, it would be rewarded if the user's simulated response[2] to that utterance included information about their emotional state. It was determined that in such implementations, training either did not converge or converged on behavior that exploited a bias of the user simulator.

   Hence, it was trialed to reward the agent's actions directly, which showed a strongly increased training effect. Different metrics to measure semantic similarity between the agent's utterance and the target were tested, including cosine similarity of normalized word embeddings using *GloVe* or *Word2Vec* akin to Tang et al.'s approach [32] and Cer et al.'s *Universal Sentence Encoder* [34]. The latter was identified to produce the most accurate and reliable results.

2. **ACHIEVING CONVERSATIONAL SMOOTHNESS: Optimization should control KL-Divergence and explore Language Model Adaptations**

   Experiments showed consistently that requiring the agent to utter specific content, such as a full target sentence, most often caused the KL-Divergence between itself and the original model to increase strongly. As illustrated by Table 3.1, the quality and relatedness of the model's generated text deteriorated commensurately with the KL-Divergence. To prevent this, it was discovered that adding Ziegler et al.'s penalty term for KL-Divergences [41] to the reward function enabled the algorithm to explore actions that approach the target behavior in closer proximity to the original model. This better prevented large deteriorations in conversation and language quality. Varying the KL-Coefficient, i.e. the weight associated with the KL-Divergence penalty, allowed better control over what range of KL-Divergences should be explored.

   However, even with this control in place, it was still difficult to maintain language quality while achieving high rates of task success. To better strike this balance, changes to the current Language Model were assessed. Whereas the current SAT-Chatbot solely uses *generation* to produce its utterances, this was extended by developing a **Mixed Generation-Retrieval** method: First, Reinforcement Learning was applied to train the agent to *generate* utterances whose

[2] User responses were simulated using a state-of-the-art chit-chat bot by Zhang et al. [54]

User Prompt: 'I just came home from football practice'

KL-Divergence	**Response Utterance**
0.0 (Baseline Model)	'Nice, What team is playing?'
10.3	'Nice! Do you still have a lot of hope?'
20.6	'I am sorry to hear that. What were you planning to do with the game?'
32.0	'Personally, I am not used to watch the news.'

Table 3.1: Examples of Chatbot Responses at different KL-Divergences

conversation topic has a high semantic similarity to the topic of 'feelings'. Crucially, a rules-based *retrieval* method was then added to the Language Model to achieve higher rates of reaching the conversational objective. Specifically, the LM computes the semantic similarity of its generated utterance to the target sequence 'How are you feeling?' and appends this sequence to the utterance if it is sufficiently semantically related[3], as illustrated in Figure 3.3. During experimentation, this approach delivered promising results, as it was able to better maintain dialogue quality whilst achieving a high task success rate.

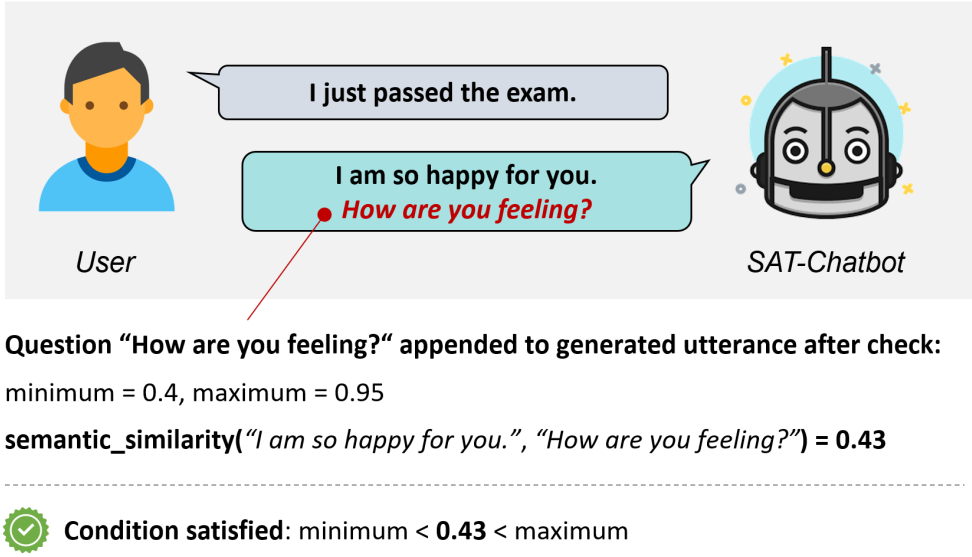


Figure 3.3: Utterance Production using Mixed Generation-Retrieval

Three additional reward signals designed to maintain and improve dialogue quality were also analyzed. First, the effect of including a reward signal for *conversational smoothness*, measured by the semantic relatedness of the agent's utterance to the user prompt using the *Universal Sentence Encoder* [34], was var-

[3] A range of values around the 80th percentile of same-utterance semantic similarity between sentences (see Section 3.2) is explored and 0.40 is identified as a good minimum cut-off for appending 'How are you feeling?'. A maximum threshold of 0.95 is also set if the LM already generated 'How are you feeling?', in which case the sequence need not be appended to the utterance again.

ied. It appeared to improve dialogue quality when training the baseline's LM design, however the effect on Mixed Generation-Retrieval models was inconsistent and warrants further exploration. Second, when rewarding the agent for including specific content in its utterances, it would tend to often explore by increasing the number of generated tokens per utterance, as this increased the probability of the utterance including the target content. Adding a *penalty for long utterance length* to the reward function effectively prevented this behavior. Third, rewarding the agent for utterances with *low perplexity* did not have the desired effect of improving the quality of English. Instead, it exploited some of the shortcomings of the metric which, e.g., appeared to assign low perplexity to utterances that repeated the same sequence of words multiple times. This echoes previous sentiment in research that this metric does not always correlate well with human judgment [28].

3. **REWARD WEIGHTING: Higher Weight should be assigned to Task Success**

   In most of the final experiments for this research, the reward signal for task success was supplemented with additional rewards for conversational smoothness. Experiments showed that assigning a higher weight to task success enabled the Language Model to converge more reliably on behavior that showed significant improvements in target-guiding over the baseline. When scaling all rewards to ranges between 0 and 1, weights of 60-70% for task success worked well. The overall training effect was not sensitive to small-scale changes in these weights.

4. **HYPERPARAMETER SETTINGS: Training effect highly sensitive to Learning Rate and KL-Coefficient**

   The learning rate and the KL-Coefficient were identified as the hyperparameters that most strongly impacted model development. They had opposite effects on training exploration: High learning rates caused the model to more quickly explore spaces with larger KL-Divergences, while high KL-Coefficients did the opposite (and vice versa). Combinations of these parameters could be used in different training strategies, e.g. using low learning rates and high KL-Divergences to explore behavior in very close proximity to the reference model.

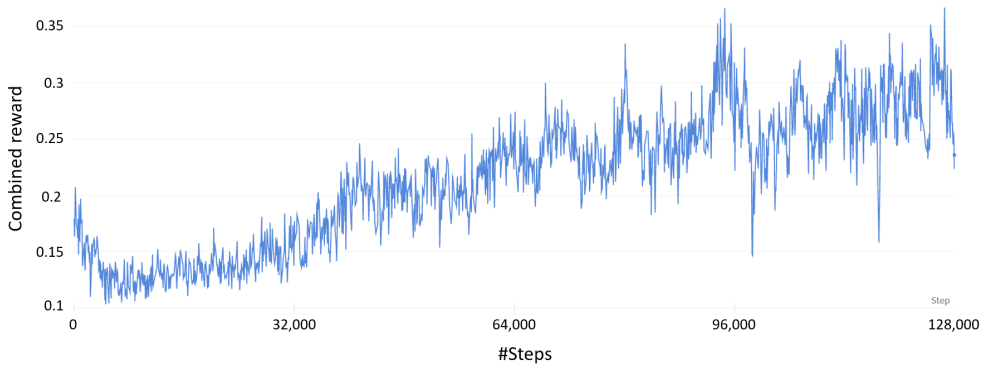


Figure 3.4: Convergence of Example Optimization Run

Regarding overall training settings, experiments showed that most training runs converged before reaching 130,000 steps, as illustrated in Figure 3.4. Crucially, when examining periodic model checkpoints during training, it was discovered that checkpoints at earlier training iterations often also offered reasonable task success rates whilst providing higher language quality. Hence, when developing the final models during this research, it is important not to solely assess the final model but also periodic checkpoints with good ratios of task success and language quality as well.

In summary, the experiments demonstrated the need to carefully choose the Language Model design and the right mix of reward signals in order to achieve a good balance of task success and conversational smoothness. The final models should not have excessively high KL-Divergences to the current baseline model and including an appropriate penalty term in the reward function enables better control of this. Overall, the training effect is most sensitive to changes in the learning rate and KL-Coefficient.

These observations are used as crucial additional input for defining the optimal Reinforcement Learning setup with Proximal Policy Optimization, as outlined in the following sections. It should be emphasized that these observations were gathered solely in the context of this project and that more fundamental research on the topic of Proximal Policy Optimization for conversational agents is strongly encouraged.

3.3.2 Reinforcement Learning Components

Previous research on RL, PPO, and dialogue systems, in addition to the lessons from the data set *DailyDialogues* and the experiments in Section 3.3.1 hint at different possible LM development approaches for achieving target-guiding. Before the chosen approaches are outlined in Section 3.4, this section first outlines the common components of their Reinforcement Learning setup:

Action

The action a undertaken by the agent is the generation of an utterance of variable length. The action space is limited only by the size of the agent's vocabulary of 50,261 tokens and a maximum utterance length of 20 tokens. This maximum utterance length was chosen as it represents the 80^{th} percentile of utterance length in a large-scale data set of human dialogue, *DailyDialogues*, with the addition of an end-of-sequence token.

State

The state or environment x is a single user utterance at a time taken from the human dialogue data set *DailyDialogues*, which is encoded by the Language Model into a fixed-size vector before feeding it through the network.

Policy

The policy π can be interpreted as the sum of the weights and biases of the Transformer Language Model, as it interprets the environment's state and produces an action. It is these weights and biases that are subject to optimization using policy gradient methods.

Rewards

Based on Ziegler et al's [41] approach, the reward R is defined as a combination of a user-defined reward r and a weighted penalty for KL-Divergences between the trained model and the baseline model from Section 3.1. The optimal weight term, the KL-Coefficient β, will be determined through hyperparameter optimization, as is explained in Section 3.6. For the user-defined reward r, each of the developed models introduced in Section 3.4 will use a subset of the following reward signals that are aimed both at eliciting the desired behavior in the SAT-Chatbot and controlling side-effects of the bot's utterance production during RL exploration. They are, among others, inspired by Gupta et al's implementation of producing single-turn topic transitions with two-sentence utterances [27].

$$R(a,x) = (\sum_{i=1}^{3} w_i \times r_i(a,x)) - \beta \times log\frac{\pi(a|x)}{\rho(a|x)}$$

where w_i is the weight associated with r_i

1) Target Proximity The aim of the first reward is to signal to the agent how close its action is to satisfying the conversational objective. Since the target of the Emotion Recognition Stage, the user revealing their emotional state, is not uttered by the agent itself, the conversational objective is reformulated in terms of the agent's actions to be able to reward it more accurately. There are two implementations of this reward signal which measure semantic similarity between the bot utterance and differing targets. Splitting these implementations has relevance for the individual models assessed in Section 3.4.

In the first implementation 1a), the bot is rewarded for producing an utterance in which the second sentence has high semantic similarity to the target sequence 'How are you feeling?'. This follows Gupta et al. by setting the second sentence of an utterance as the target sentence [27], leaving flexibility in the generation of the *bridging sentence*, the first sentence of the utterance.

Algorithm 1 - Reward Signal: Target Proximity-Implementation 1a

$bot\_utterance = a$
$second\_sentence = bot\_utterance[1]$
$target =$ 'How are you feeling?'
r_1 = semantic_similarity($second\_sentence$, $target$)

In the second implementation 1b), the bot is rewarded for choosing conversation topics in close proximity to the target topic of 'feelings'. This is implemented by using a high-performing keyword extraction algorithm developed by Campos et al. [55] to identify the keywords within each sentence of a bot utterance, compute the semantic similarity between each of the keywords and the topic 'feelings', and rewarding the agent with the highest similarity score. This implementation is akin to the keyword-based approach to target-guiding proposed by Tang et al [32].

Algorithm 2 - Reward Signal: Target Proximity-Implementation 1b

$bot\_utterance = a$
$target =$ 'feelings'
for each $sentence$ **in** $bot\_utterance$ **do**
    $keyword$ = extract_keyword($sentence$)
    $similarity$ = semantic_similarity($keyword$, $target$)
    if $similarity > best\_similarity$ **then**
        $best\_similarity = similarity$
    end if
end for
$r_1 = best\_similarity$

As a measure of semantic similarity for implementations 1a) and 1b), Cer et al's [34] approach is used to compute the inner product between the encodings of a bot-generated sentence/keywords and the target 'How are you feeling?'/'feelings', employing the *Universal Sentence Encoder*. It is scaled between 0 and 1, with high numbers representing high semantic similarity.

2) Conversational Smoothness This reward signal is designed to incentivize the agent to produce an utterance that is highly related to the user prompt and fits the flow of the conversation. It is inspired by Faal et al's implementation that optimized for the degree to which a prompt and bot response are related by information [42]. Again, semantic similarity using the *Universal Sentence Encoder* is employed to reward conversational smoothness in two ways: First, the semantic similarity between the user prompt and the first sentence of the utterance produced by the bot is measured. This is intended to increase the relevance of the response to the user. Second, the average semantic similarity between all sentences of the generated bot utterance is computed, in an effort to prevent abrupt topic switches over the course of the bot utterance, which the agent might otherwise explore to satisfy the conversational target of the reward signal 1). Separating incentivization this way, instead of simply rewarding similarity between the user utterance and the entire bot utterance, allows the agent to optimize the relatedness of its first sentence separately to its other sentences. This provides more flexibility in utterance optimization in order to produce a *bridging sentence* akin to Gupta et al. [27]. Again, both measures are scaled between 0 and 1, with high numbers representing high semantic similarity. They are weighted to produce a single scalar reward for conversational smoothness, applying a higher weight of $w_u = 70\%$ to the relatedness between the user prompt and the

first bot sentence. This is intended to disincentivize the bot from exploring actions that repeat overly similar sentences over the course of its utterance.

Algorithm 3 - Reward Signal: Conversational Smoothness

$user\_prompt = x$
$bot\_utterance = a$
$first\_bot\_sentence = bot\_utterance[0]$
$w_u = 0.7$
$user\_bot\_relatedness$ = semantic_similarity($user\_prompt, first\_bot\_sentence$)
for each $sentence$ **in** $bot\_utterance$ **do**
    $similarity$ = semantic_similarity($sentence, sentence + 1$)
    $sentence\_similarity\_list$.add($similarity$)
end for
$within\_bot\_relatedness$ = average($sentence\_similarity\_list$)
$r_2 = w_u \times user\_bot\_relatedness + (1 - w_u) \times within\_bot\_relatedness$

3) Controlling for Utterance Length Based on the insights of Section 3.3.1, training the agent solely with reward signals 1) Target Proximity and 2) Conversational Smoothness will cause the bot to explore long utterances with many sentences. At the same time, bot utterances should not be comprised of too few sentences, as this would adversely affect their ability to relate to the user prompt and include the conversational target. Gupta et al's objective of optimizing for an utterance length of two sentences [27] is followed, introducing a penalty of 0.1 for every sentence generated too many or too few.

Algorithm 4 - Reward Signal: Utterance Length

$bot\_utterance = a$
$sentence\_count$ = length($bot\_utterance$)
$target\_sentence\_count = 2$
$penalty = -0.1$
$delta$ = abs($sentence\_count$ - $target\_sentence\_count$)
$r_3 = delta \times penalty$

Logic of Optimization

Algorithm 5 illustrates the Reinforcement Learning training loop. For every optimization step, the agent is provided with a single user prompt and generates a response utterance as per its policy. Both the user-defined reward r and the weighted KL-Divergences between the agent's policy π and the baseline's policy ρ are then computed, and the combined reward R is used for optimization. Whereas the user-defined reward scheme and this training setup incentivizes the models to perform *single-turn transitions*, the inclusion of the KL-penalty is expected to balance this and cause the agents to learn to progressively push the conversation more into the direction of 'feelings'. If learned successfully, it is expected that the Language Models effectively execute a *multi-turn* conversation strategy as a result.

Algorithm 5 - PPO Optimization Loop

$\pi = baseline\_sat\_model$
$\rho = baseline\_sat\_model$
$\beta = KL\_coefficient$, $\alpha = learning\_rate$
for each $epoch$ **do**
  x = retrieve_user_prompts($DailyDialogues$)
  a = π.respond(x)
  $KL = \log \frac{\pi(a|x)}{\rho(a|x)}$
  $R = (\sum_{i=1}^{3} w_i \times r_i(a, x)) - \beta \times KL$
  $loss = ppo\_optimizer$.loss(π, R)
  $ppo\_optimizer$.step($\pi, loss, \alpha$)
end for

3.4 Developed Language Models

Based on the aforementioned previous research on Reinforcement Learning and Proximal Policy Optimization for NLP, as well as the preliminary experiments, three promising approaches are defined to achieve the objective of making the SAT-Chatbot target-guided. The choice of models reflects observations of the potential that adaptations to the Language Model and reward signals have for balancing task success with conversational smoothness. First, a **Pure Generative** design is explored which trains the baseline without changes to the utterance production strategy, yielding Model **A**. Finally, two variants of **Mixed Generation-Retrieval** models are developed, which feature different reward schemes to produce Models **B1** and **B2**.

Pure Generative Model

Model A The first model relies on Reinforcement Learning end-to-end to fine-tune the GPT-2 model outlined in Section 3.1 for target-guiding behavior. In addition to optimizing for 2) Conversational Smoothness and 3) Controlling for Utterance Length, target reward signal 1a) is used to incentivize the agent to produce utterances that include a sentence with close semantic proximity to 'How are you feeling?'. Following experimentation, weights for the individual reward signals are identified, arriving at the following composite reward function:

$$R(a,x)_A = 0.65 \times r_{1a}(a,x) + 0.35 \times r_2(a,x) + r_3(a,x) - \beta \times log\frac{\pi(a|x)}{\rho(a|x)}$$

This model is included following the expectation that a well-developed, purely generative Language Model produces flexible dialogue and paths toward the target utterance. Simultaneously, the challenging nature of this optimization especially with regards to maintaining a good quality of English should be noted, as this cannot be explicitly controlled for in the reward function.

Mixed Generative-Retrieval Models

Here, Reinforcement Learning is first applied to train the dialogue policy of Language Models **B1** and **B2** to choose conversation topics that guide towards the target. Crucially, the LM is then further supported by a rules-based retrieval method to achieve higher rates of reaching the conversational objective. The rules mirror those explained in Observation 2 of Section 3.3.1, in that the question 'How are you feeling?' is appended if its semantic similarity with the *generated* bot utterance is higher than 0.40 and below 0.95.

The Reinforcement Learning setup for Language Models **B1** and **B2** differs in the combinations of reward signals employed:

Model B1 The Transformer in this Language Model is trained using a combination of all three reward signals, similarly to Model **A**. Importantly, reward signal 1b) is chosen as the incentive for task success, as the objective is only to steer the conversation in the direction of the conversational topic 'feelings' and not to *generate* the target utterance 'How are you feeling?'. Reward signals 2) and 3) are used to incentivize the model to generate relevant two-sentence utterances. The final composite reward function is as follows:

$$R(a,x)_{B1} = 0.65 \times r_{1b}(a,x) + 0.35 \times r_2(a,x) + r_3(a,x) - \beta \times log\frac{\pi(a|x)}{\rho(a|x)}$$

The expectation for optimizing for a conversational *topic* instead of a full sentence is to provide the PPO optimization with more flexibility in exploring paths to the target. Since the question 'How are you feeling?' is not *generated* but *appended*, the Transformer will likely maintain a higher quality of English compared to Model **A**.

Model B2 This model is nearly identical to Model **B1**, however, it only trains the Transformer in the Language Model to produce 1b) on-target and 3) two-sentence utterances. It omits reward signal 2) Conversational Smoothness to yield the following composite reward function:

$$R(a,x)_{B2} = r_{1b}(a,x) + r_3(a,x) - \beta \times log\frac{\pi(a|x)}{\rho(a|x)}$$

Model **B2** follows the observation from experimentation that Mixed Generation-Retrieval models did not always require a dedicated reward signal for conversational smoothness to maintain dialogue quality (see Observation 2). This model is expected to marginally outperform Model **B1** in terms of task success and produce at least comparable conversational smoothness.

3.5 Evaluation

The assessment of these models is designed to measure the extent to which they are target-guiding, i.e. whether their conversations reach the objective of the user revealing their emotional state smoothly. Additionally, it should also be captured whether previous behavior that the Transformer model was fine-tuned for, i.e. engaging in empathetic dialogue, is still present. To fulfill these objectives, an automatic assessment, where bot-user interactions are simulated, is combined with a non-clinical human trial.

The often poor conversational quality of the baseline model as outlined in Section 3.1 highlights the need to assess the performance of the optimized models *relative* to this baseline[4] and the assessment is designed explicitly to maximize comparability between all models. Model **A** is evaluated against the original **Pure Generation** baseline Transformer model delineated in Section 3.1. To interpret the performance of Models **B1** and **B2**, a **Mixed-Generation-Retrieval** baseline model is constructed by adding the rules-based retrieval method for appending the question 'How are you feeling?' to the baseline from Section 3.1.

Automized Assessment

An automized assessment was included since testing the behavior of a conversational agent over a large number of conversations is highly labor-intensive. Here, simulated chat interactions between the Language Models and a leading chit-chat bot are simulated, which follows similar approaches introduced in Section 2.3. For the chit-chat bot, Microsoft's *DialoGPT Medium* Language Model is chosen, a GPT-2 based conversational agent that has demonstrated response quality comparable to humans under a Turing Test and produces more related and contentful responses compared to other Language Models [54]. Scrutiny of simulated chat transcripts reveals that the far majority of chats mimic the conversation quality of the human trial. *DailyDialogues* data as outlined in Section 3.2 is employed to source conversation starters, i.e. the first utterances of the 1,312 conversations in the test set. These are used as the first user utterance in the simulated chat - the assessed Language Model then responds and the chit-chat bot and the LM take turns responding to one another. Each chat simulation lasts until the target of the conversation or a maximum of 20 turns is reached, whichever comes first.

To capture whether the target of the conversation, which is the user revealing their emotional state, is reached, the simulated user utterances are analyzed for uses of keywords that signal emotions, e.g. 'happy', 'anxious', and 'sad'. These keywords are selected from Encyclopædia Britannica's larger 'Emotions Vocabulary' list [56]. This approach was found to perform more reliably for this purpose compared to existing emotion classifiers developed in the context of SAT [6; 7], which were designed to

[4] Whereas Yan's implementation of the SAT-Chatbot forcefully poses the question 'How are you feeling?' after 7 turns [7], this provision is removed for evaluation in order to be able to derive more general conclusions on the effect of PPO training for target-guiding.

classify *what kind* of emotion is uttered, not to detect *whether* an emotion was uttered. Crucially, measuring the conversational objective this way, by analyzing *user* responses, represents a difference to how the models are trained: During training, the *agent* is rewarded for choosing utterances in close proximity to 'How are you feeling?'/the conversational topic 'feelings'. During evaluation, analyzing *user* utterances for task success is strongly aligned with capturing the conversational goal of the SAT-Chatbot's Emotion Recognition Stage.

Across the simulated conversations between the Language Models and the user simulation chit-chat bot, five metrics are captured in an automized fashion:

- **Task Success Rate**: Share of conversations that reached their conversational objective of the user revealing their emotional state within 20 turns. This is a measure of the *effectiveness* of the target-guiding behavior, and the objective is to increase it compared to the baseline.
- **Turns to Target**: Within the set of conversations where the objective was reached, the average number of turns until the target was achieved. This is a measure of the *efficiency* of the target-guiding behavior, and the objective is to decrease the Turns to Target compared to the baseline.
- **User-Bot Relatedness**: Semantic similarity between the user utterance and the first sentence of the bot utterance, as captured by the *Universal Sentence Encoding* approach from Section 3.3.2. This is a measure of the *response-relatedness* of the dialogue, and the objective is to maintain it compared to the baseline.
- **Within-Bot Relatedness**: Average semantic similarity between the sentences of the bot utterance, as captured by the *Universal Sentence Encoding* approach from Section 3.3.2. This is a measure of the *continued fluency of the dialogue*, and the objective is to maintain it compared to the baseline.
- **Perplexity**: Perplexity of the generated bot utterance. This is a measure of the extent to which the Language Model has *preserved its generative capabilities*. As perplexity has a large standard deviation even when comparing different utterances of the same Language Model, the objective is only to prevent it from increasing too strongly compared to the baseline.

Human Trial

A human trial is included since, as mentioned in Section 2.3, the assessment of language quality is often not captured well by automized metrics and can be subjective [28]. The participants chat with each of the three trained models, in addition to the two baseline models, and are not informed of the conversational objective. Each conversation starts with the participants answering the same question by the bot 'What did you do today?'. The chat interaction continues for a maximum of 20 turns or until the target of the conversation is reached by the user revealing their emotions. After each completed chat, participants fill out a questionnaire to answer the same set of questions for each of their chat interactions, which are designed to

capture the extent to which the conversation is target-oriented, smooth, and empathetic. This questionnaire can be found in Appendix B. Respondents score their agreement to a statement using a 5-point Likert Scale, which during analysis is averaged by assigning a score of 1-5 to the responses *Strongly disagree* to *Strongly agree*, respectively. The chat transcripts of the trial participants with the models are analyzed separately to determine the target-guiding's effectiveness, measured by the aforementioned *Task Success Rate*, and its efficiency, as measured by the *Turns to Target*. The Language Models are made available to the participants using Google Colab, and Google Forms is used to capture questionnaire responses.

3.6 Training Settings and Hyperparameters

This section elaborates on the process by which the final training settings are set, optimal hyperparameters are defined, and promising model checkpoints are identified.

Regarding training settings, the implementation is based on Ziegler et al. by using the Adam optimizer, performing 4 PPO epochs per batch, and setting the clip value for policy updates to 0.2 [41]. The discount factor for the estimated value function, γ, is defined as 0.99 to incentivize near-term exploration of high-reward policy changes. Following experimentation as mentioned in Section 3.3.1, 128,000 steps per training are performed in batches of 64.

For each of the models **A**, **B1** and **B2**, performance is assessed with the following hyperparameters, the range of which is, again, inspired by previous research [41]:

- **Learning Rate**: $[1.01 \times 10^{-5}; 1.21 \times 10^{-5}; 1.41 \times 10^{-5}]$
- **KL-Coefficient** β: $[0.01, 0.15, 0.3]$

As illustrated in Figure 3.5, all combinations of model types and training settings enable the PPO algorithm to optimize performance.

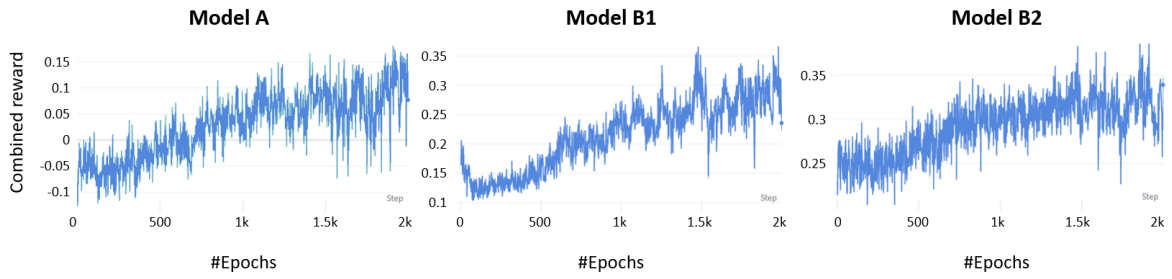


Figure 3.5: Rewards during Example Training Runs of Models **A**, **B1** and **B2**

In order to identify the optimal hyperparameter settings, a total of 27 models (3 Model Types **A, B1** and **B2** $\times$ 3 Learning Rates $\times$ 3 KL-Coefficients) are trained and the performance of periodic training checkpoints[5] of these models is assessed using

[5]Per model type, on average 4-6 checkpoints are assessed that showed a good ratio of user-defined reward to KL-Divergence. Across all model types and training settings, a total of 140 model checkpoints are assessed.

the user simulations and metrics of the automized assessment scheme defined in Section 3.5, run on the validation set of conversation starters.

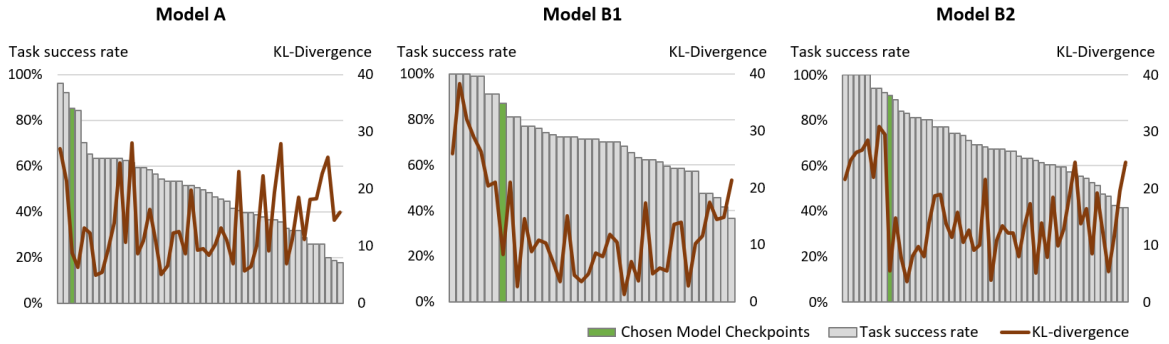


Figure 3.6: Validation Performance and KL-Divergence of Model Checkpoints

In line with the objective of this research, i.e. achieving the LM's conversational objective smoothly, the final training configuration per model type is selected by ranking all assessed model checkpoints by Task Success Rate and choosing the first model with a KL-Divergence below 10.0. Figure 3.6 displays the performance of the range of assessed model checkpoints on the validation test set. The chosen models are highlighted in green and their training settings are provided in Table 3.2. Their performance is examined in greater detail in the evaluation of Section 4.1 and Section 4.2.

	Selected Model			Val.-Performance	
Model	**Learning Rate**	**KL-Coeff.**	**Checkpoint**	**Task Success Rate**	**KL-Div.**
A	1.21×10^{-5}	0.01	10	85.1%	8.8
B1	1.41×10^{-5}	0.30	306	87.1%	8.2
B2	1.01×10^{-5}	0.30	220	91.1%	5.5

Table 3.2: Optimal Model Settings after Hyperparameter Search

3.7 Ethical Considerations

An *Ethics Checklist* provided by Imperial College London [57] is used to guide the discussion of the ethics of this research. In the following, the human trial and provisions for the use of the final SAT-Chatbot are addressed.

As outlined in Section 3.5, the evaluation includes a non-clinical human trial in which the participants engage with different versions of the chatbot and provide feedback on their interaction. Especially the fact that they interact freely with the chatbots and respond to the question 'How are you feeling?' calls for the protection of their personal information. This is addressed specifically by anonymizing all recorded information, i.e. questionnaire responses and chat transcripts, and by linking responses solely through a *Survey-ID* generated by the Google Colab platform that is used to distribute the study. This protects the identity of the individuals

whilst allowing to link chat interactions with questionnaire responses, which is required for evaluation. As a final note on this point, it should be mentioned that trial participants' answers to the question 'How are you feeling?' in chit-chat interactions produced generally-formulated responses (e.g. 'I feel okay', 'I am feeling fine') that are not very intrusive.

As this research project continues the development of the SAT-Chatbot for the eventual application in psychotherapy, it is suggested that additional safety measures be implemented before testing the chatbot on depressed patients. The dialogue generated by the GPT-2 Transformer currently flexibly responds to user inputs which, without provisions, carries the risk of generating responses that make the user feel neglected or underappreciated [12]. Hence, it is recommended to add measures to the SAT-Chatbot's Language Model which ensure, e.g., that sensitive topics such as self-harm are not encouraged, and that if any inclination towards self-harm or even suicide were detected, the bot is able to provide highly relevant and immediate support (e.g. providing the user with the telephone number of a suicide hotline).

Chapter 4

Results and Discussion

As outlined in Section 3.5, models' performance is evaluated by combining an automized assessment with a non-clinical human trial. This human trial was conducted with 22 participants, with an equal share of 50% male and female respondents predominantly in the age group of 25-40 years.

Our evaluation identifies Model **B2** as clearly outperforming the baselines and other RL-trained models both in terms of the effectiveness of reaching the conversational objective and dialogue smoothness. In this chapter, first the performance increase of Model **B2** compared to the current SAT-Chatbot is reported in Section 4.1. Thereafter, a more differentiated analysis of the performance of each of the three RL-trained models is performed. The insights from these two sections will inform the discussion of the training effect of PPO on dialogue systems in Section 4.3.

4.1 Principal Findings

Model **B2**, which follows a Mixed Generation-Retrieval approach and was trained with Proximal Policy Optimization primarily for target proximity, outperforms the baselines as well as other training approaches by a wide margin. The following describes separately how **B2** is highly effective in reaching the conversational objective and how it displays positive side-effects on dialogue quality.

4.1.1 Task Success

Analyzing the effectiveness of the models' dialogue strategy, a significant improvement in the *reliability* to which the optimized Model **B2** reaches the target of the conversation, having the user reveal their emotional state, can be observed. An improvement in the *efficiency* with which the target is reached is also demonstrated.

Figure 4.1 shows the *Task Success Rate*, i.e. the share of conversations in which the target was reached within 20 turns. It measures the *reliability* to which the conversational objective is achieved. Model **B2**, which achieves its target in ˜82-

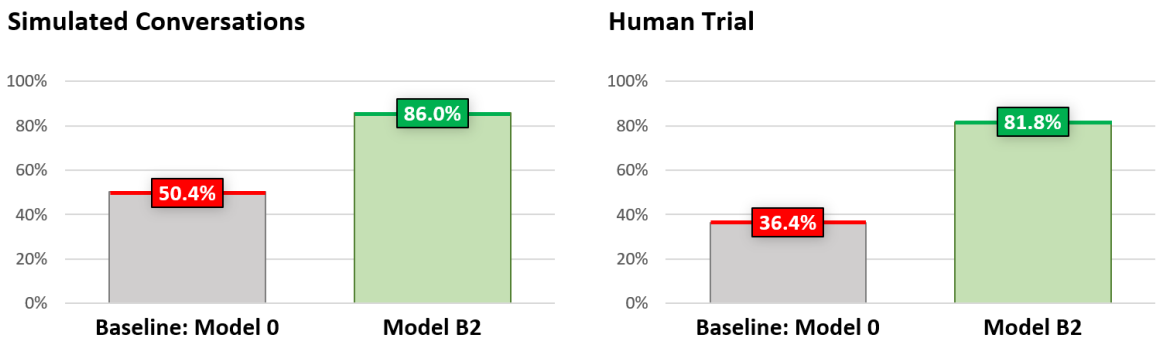


Figure 4.1: Task Success Rate of Model 0 and Model B2

86% of dialogues, outperforms the baseline by wide margin of ˜$36 - 45\%$pts. This effect appears comparable when analyzing both the simulated conversations and the human-to-bot interactions of the non-clinical trial.

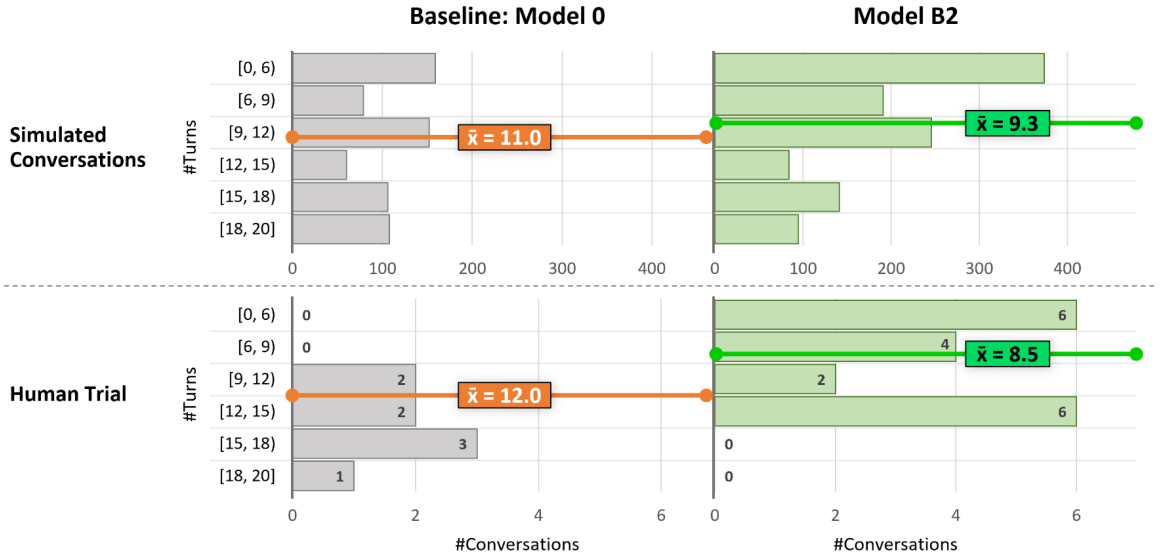


Figure 4.2: Number of Turns until Target reached in Successful Conversations

Additional analysis is performed on dialogues in which the conversational objective was achieved, where the number of turns until the user uttered the target is counted - a measure of the *efficiency* of the models' target-guiding. Figure 4.2 once again shows a significant performance increase of the optimized Model **B2**, which achieves the target in 16-30% fewer turns compared to the baseline. The results are, once again, consistent across simulated and human-to-bot interactions. An analysis of the distribution of the results demonstrates that Model **B2** achieves its performance increase especially through shifts at the extreme ends of the turn count spectrum. This effect is strongest in the human trial: Whereas 50% of successful interactions with the baseline had > 15 turns, Model **B2** reduced this share to 0%. At the same time, 33% of Model **B2**'s successful chats were comprised of < 6 turns which stands in stark contrast to the baseline's most efficient interaction, which consisted of 10 turns. The automized assessment shows similar shifts.

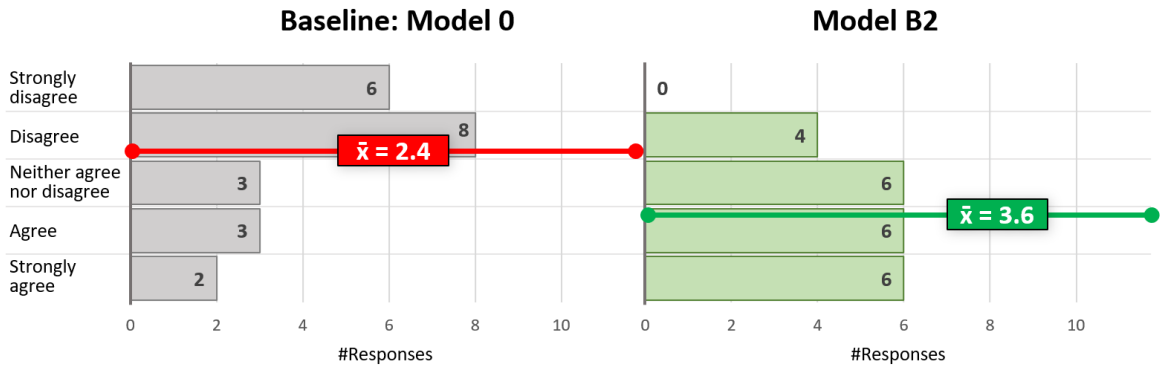


Figure 4.3: Human Trial Participants' agreement with statement: '**I felt that the Chatbot was interested in how I was feeling**'

Whereas the previous two evaluations were based on analyses of the chat transcripts, the participants of the human trial concurred that Model **B2** led the conversation in the direction of the user's feelings. Figure 4.3 shows that, on average, the respondents *agreed* with the statement 'I felt that the Chatbot was interested in how I was feeling'. Simultaneously, respondents on average *disagreed* with the statement for the baseline, a symptom of its design for chit-chat without elements of target-guiding.

4.1.2 Conversational Smoothness

Compared to the baseline, the evaluation of Model **B2**'s conversational smoothness reveals a significant increase in the relatedness, engagement, and empathy of the dialogue.

Baseline: Model 0

Chatbot:	What did you do today?
User:	I had a long sleep
Chatbot:	I am sorry to hear that. What happened.
User:	I was really tired
Chatbot:	Was the work done?
User:	I am not sure what you mean
Chatbot:	Did you go to the office and pick it up.
User:	No, I am working from home
Chatbot:	Well then I hope you do go!
[...]	

Model B2

Chatbot:	What did you do today?
User:	I had a long sleep
Chatbot:	Oh well, You should feel better! How are you feeling?
User:	I'm upset, I still feel tired unfortunately

Figure 4.4: Human Trial Transcript - Model **0** and **B2**'s Responses to same initial Prompt

Our human trial participants chatted with all models and immediately noted the often poor conversational quality of the baseline model, as first introduced in Section 3.1, stating that its responses were predominantly not relevant and felt unnatural. Model **B2**'s interactions were consistently ranked higher in terms of dialogue quality. Example chat transcripts from the Human Trial that illustrate this performance increase are found in Figure 4.4.

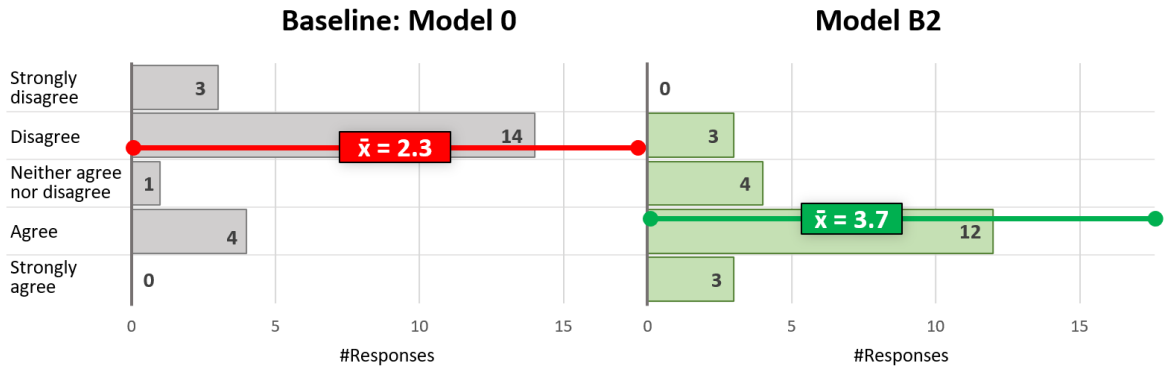


Figure 4.5: Human Trial Participants' agreement with statement: '**The Chatbot's Responses were relevant to what I was saying**'

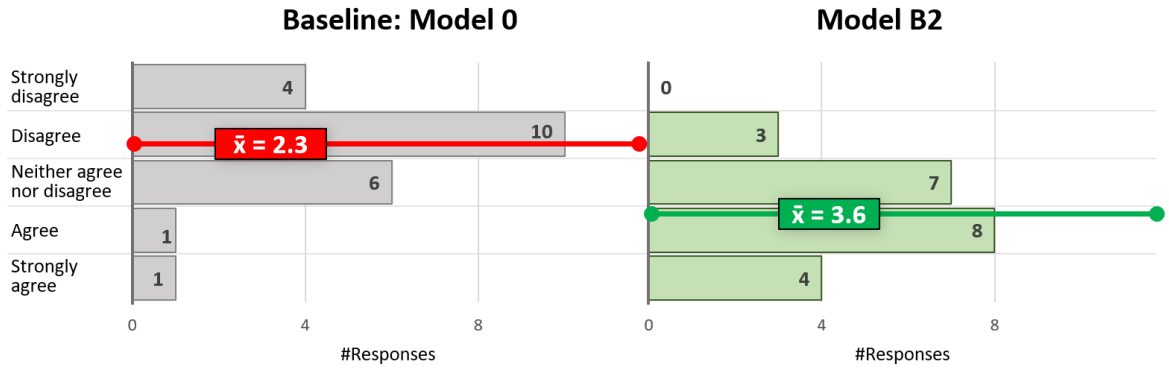


Figure 4.6: Human Trial Participants' agreement with statement: '**I had an engaging and natural Conversation with the Chatbot**'

After chatting with the models, trial participants were asked to gauge whether the chatbots' responses had high *relevance* to the conversation. Analyzing their responses provides insight into the extent to which the content of a bot utterance is semantically related to the content of a previous user utterance. Figure 4.5 underscores the previously mentioned often poor performance of the baseline, with 77% of participants either *disagreeing* or *strongly disagreeing* with the statement. This number decreases sharply to 14% for Model **B2** and participants on average *agree* with the statement, signifying a strong improvement of the PPO-trained model relative to the baseline.

A similar pattern emerges when analyzing whether participants felt the conversation was engaging, as shown in Figure 4.6. This measures the extent to which the bot maintained a natural conversation, e.g. by engaging the user with questions instead of giving short apathetic answers. It confirms earlier observations on dialogue quality both in terms of the poor dialogue quality of the baseline and a performance increase of similar magnitude with Model **B2**.

Finally, in Figure 4.7, trial participants' assessment of whether the models displayed empathy during dialogue is analyzed. As explained in Section 3.1, the baseline model was explicitly trained for this, and its performance on this statement is an

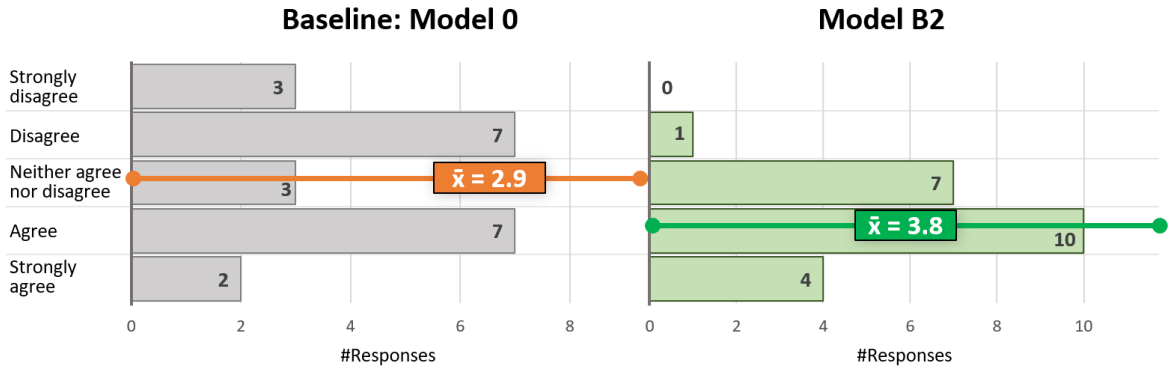


Figure 4.7: Human Trial Participants' agreement with statement: '**The Chatbot's Responses were empathetic**'

improvement on its previous scores in Figures 4.5 and 4.6. The degree to which it demonstrates empathy is, however, still inconsistent, with respondents on average *neither agreeing nor disagreeing* with the statement. Again, Model **B2** is able to display a performance improvement relative to the baseline, with respondents on average *agreeing* and only 5% of participants *disagreeing* that the dialogue was empathetic.

4.2 Differentiated Analysis of Training Approaches

As introduced in Section 3.4, three different LM development approaches were explored, reflecting a combination of reward signals and Language Model designs. This section analyzes the performance of each of the PPO-trained Models **A**, **B1** and **B2** between each other. As outlined in Section 3.5, performance is compared to two baseline models:

- Model **0**: Previous version of the SAT-Chatbot, as introduced in Section 3.1
- Model **0-Mix**: Mixed Generation-Retrieval implementation of Model **0**

Including Model **0-Mix** as a secondary benchmark will aid the evaluation of the task success of the Mixed Generation-Retrieval Models **B1** and **B2**, since it enables differentiating between performance increases attributable to Language Model design and to PPO training.

This section demonstrates the superior performance of Model **B2** compared to alternative LM development approaches and the baselines in the far majority of metrics to assess task success and conversational smoothness. Again, these two categories are analyzed separately and their performance is shown in all metrics measured across the automized assessment and human trial. For the automized assessment, additional Two-Sided *t*-tests are performed to determine whether the difference in the means of the models' results is significant. Insights from this section will contribute to the discussion on the success of PPO approaches for target-guiding in conversational agents in Section 4.3.

4.2.1 Task Success

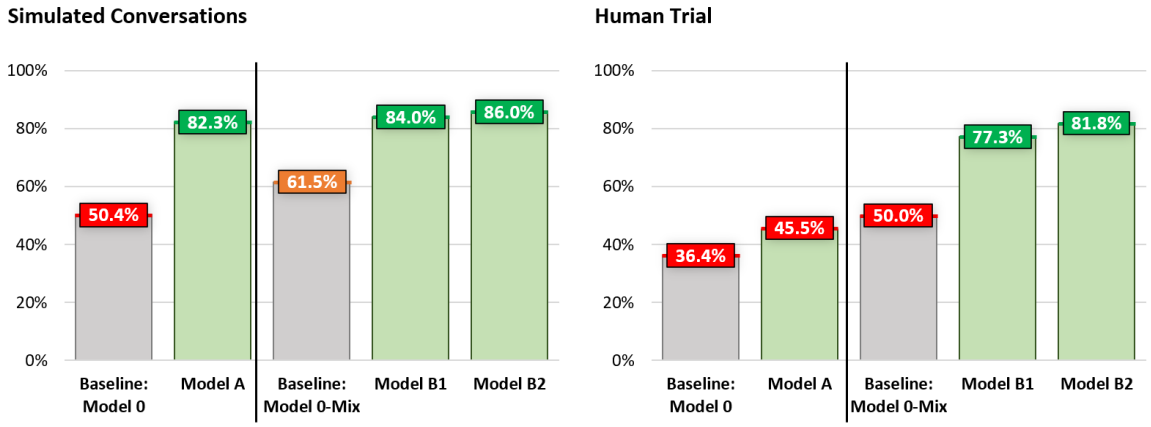


Figure 4.8: Task Success Rate across Models

Figure 4.8 displays the proportion of conversations between the SAT-Chatbot and the user simulator that achieved the conversational goal. It shows clearly that RL-trained models, especially those following a Mixed Generation-Retrieval strategy, outperform their baselines. This effect is most consistent for Models **B1** and **B2**, which outperform their Baseline Model **0-Mix** with margins of ˜30%pts. Out of the two, Model **B2** reaches the target more reliably by a narrow margin compared to **B1**. Model **0-Mix**, which appends the question 'How are you feeling?' to the end of an utterance if it is sufficiently semantically related, already outperforms Model **0** by a small margin. Whereas this indicates that the Mixed Generation-Retrieval strategy for utterance production already slightly increases the Task Success Rate, training these models with PPO and thereby yielding **B1** and **B2** still amplifies this effect by a wide margin. Model **A** initially follows the same pattern of outperforming its baseline in automized assessment, however this effect is not confirmed in the human trial. Here, this model only achieves the conversational target in +9%pts of dialogues compared to the Baseline Model **0**.

For the set of conversations that reached the conversational target, Figure 4.9 displays the average number of turns each model required until the objective was achieved. Table 4.1 provides supporting information by showing whether the differences in mean performance are statistically significant. The human trial in particular displays that all trained models reach the conversational target more efficiently compared to Baseline Model **0**. Importantly, the same results also point to a general effect that the Mixed Generation-Retrieval strategy has on target-guiding, with Model **0-Mix** performing comparably to the trained models. Model **B1** performs most efficiently in the human trial by a wider margin, reaching the target in 6.6 turns, 45% quicker than Model **0**. It also outperforms other Mixed Generation-Retrieval models by 1.4-1.9 turns. Model **A**, **0-Mix** and **B2** perform similarly, within a narrow band of 0.5 turns. The automized assessment displays a similar pattern of all trained models outperforming the Baseline Model **0**. Whereas this gain in efficiency is statistically significant at p=0.01, it is, however, not as high in magnitude compared to the human trial. Specifically, the increase in efficiency over Model **0** is ˜18% at maximum,

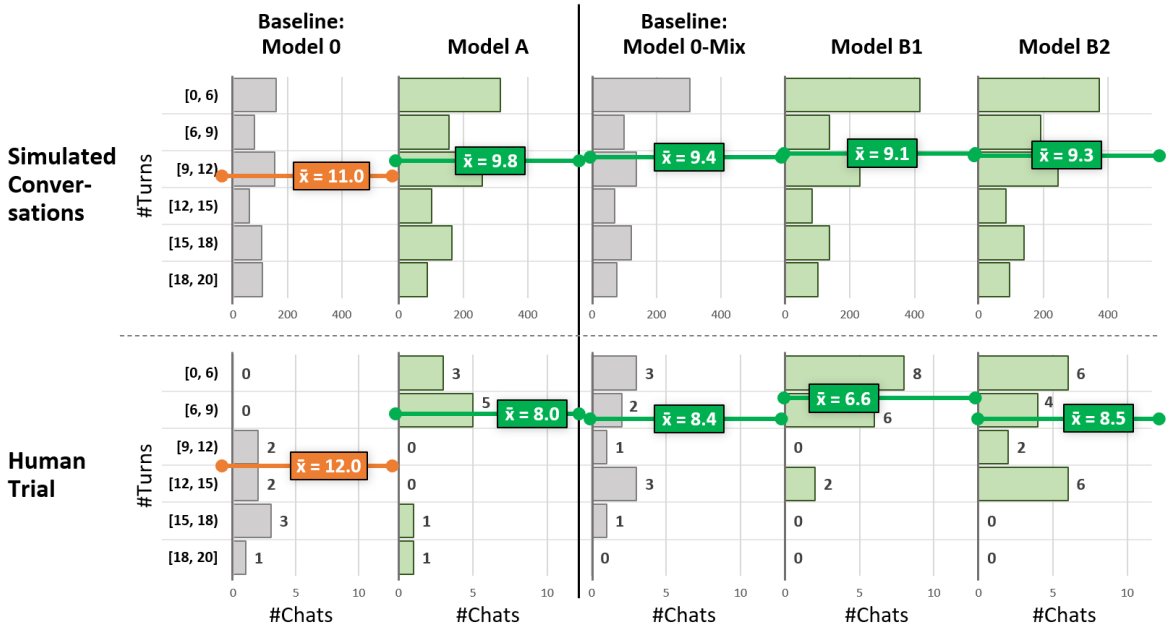


Figure 4.9: Number of Turns until Target reached in Successful Conversations

and also the strong outperformance of Model **B1** as observed in the human trial is more muted. No statistically significant performance differences in task efficiency can be observed amongst the set of Mixed Generation-Retrieval models, including their Baseline Model **0-Mix**.

	Model **0**	Model **A**	Model **0-Mix**	Model **B1**	Model **B2**
Model **0**					
Model **A**	**0.00**				
Model **0-Mix**	**0.00**	0.17			
Model **B1**	**0.00**	**0.00**	0.16		
Model **B2**	**0.00**	0.02	0.48	0.42	

Table 4.1: p-Value Results of Two-Sided t-Test on Number of Turns until Target (Automized Assessment)

As displayed in Figure 4.10, the participants of the trial agreed that especially Models **B1** and **B2** were interested in their feelings. This result shows parallels with the overall rates of task success analyzed in Figure 4.8. Mixed Generation-Retrieval models trained with PPO outperform other models, including the overall Baseline Model **0** by a large margin, with Model **B2** performing highest.

Summary of Comparative Task Success

Summarizing the comparative analysis of the target-driven behavior of the models, Model **B2** displays a strong and largely consistent superior performance compared to other models, both across the automized assessment and human trial. Mixed Generation-Retrieval models in general also outperformed the Baseline Model **0**,

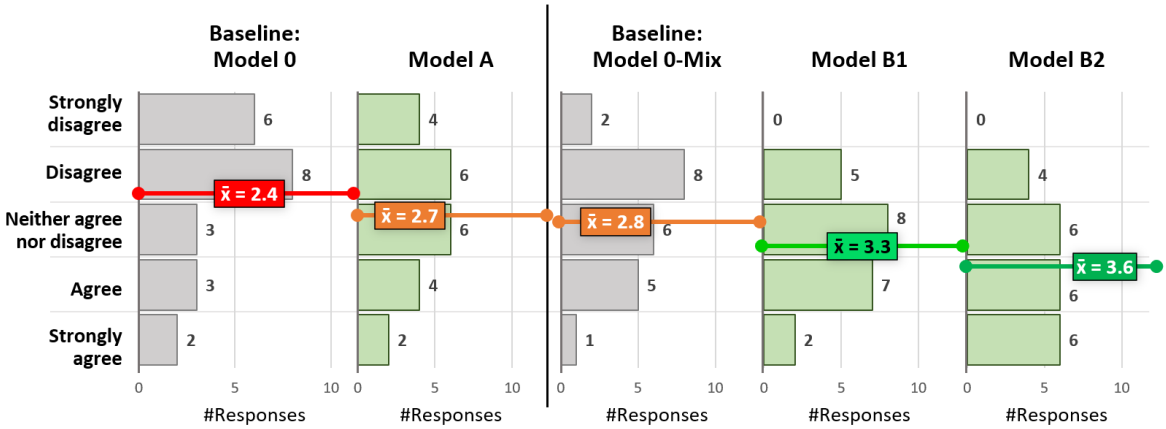


Figure 4.10: Participants' agreement with statement: '**I felt that the Chatbot was interested in how I was feeling**'

though training with PPO strongly amplified the target-driven behavior, especially regarding the reliability at which the target is achieved. The Generation-only Model **A** yielded mixed results: Whereas it performed similarly to Models **B1** and **B2** in the automized assessment, its performance fell short in the human trial.

4.2.2 Conversational Smoothness

Again, the comparative analysis of conversational smoothness in terms of dialogue quality, level of English, and empathy is focused mainly on insights from the human trial. As supporting evidence, additional analysis on the *User-Bot Relatedness*, *Within-Bot Relatedness*, and the *Perplexity of the Bot Utterance* during the automized assessment is provided.

Dialogue Quality

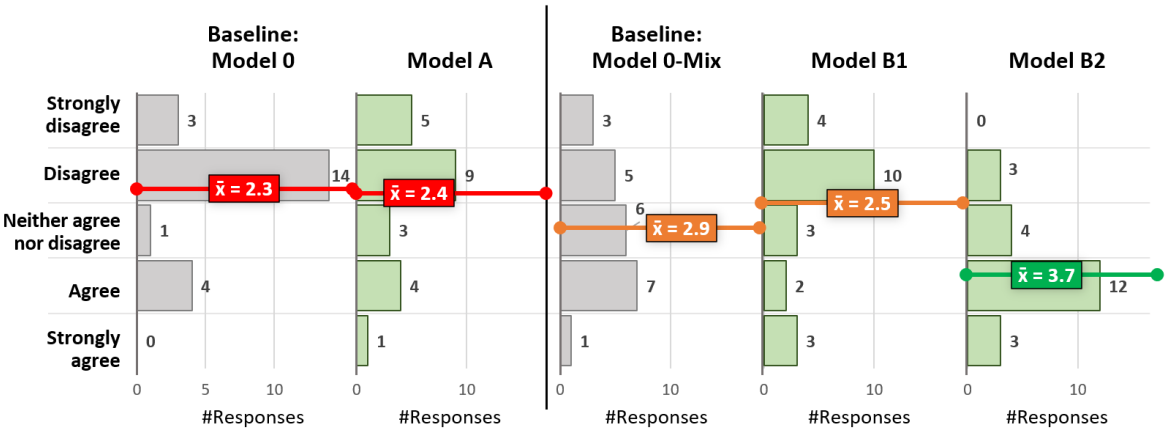


Figure 4.11: Participants' agreement with statement: '**The Chatbot's Responses were relevant to what I was saying**'

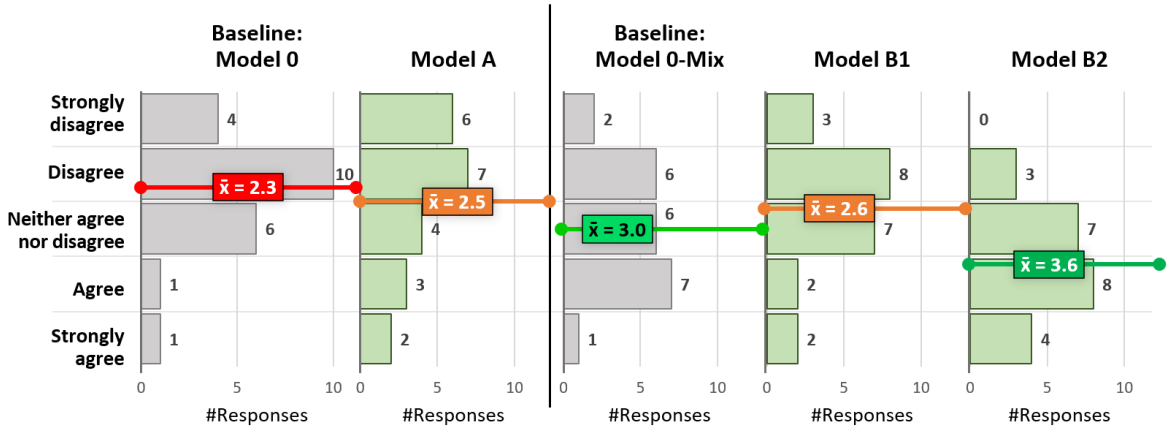


Figure 4.12: Participants' agreement with statement: '**I had an engaging and natural Conversation with the Chatbot**'

Figure 4.11 shows that the positive side-effect target-driven RL-training can have on conversational quality, as shown for Model **B2** in Section 4.1, is not consistent across all model training approaches. Instead, Models **0**, **A**, **0-Mix** and **B1** perform roughly comparably, with a slight penalty for Generation-only models. In Models **A** and **B1**, ˜63% of respondents still disagreed or strongly disagreed that the bots' responses were relevant. Despite not having been explicitly trained for conversational smoothness, Model **B2** is the only model to outperform strongly and reliably.

A similar pattern is also visible in participants' assessment of whether dialogue was engaging, as shown in Figure 4.12. As an additional note, Model **0-Mix** already outperforms the Baseline Model **0**, seemingly only through the addition of the question 'How are you feeling?'. Also in this analysis, however, Model **B2** comparatively outperforms other models overall.

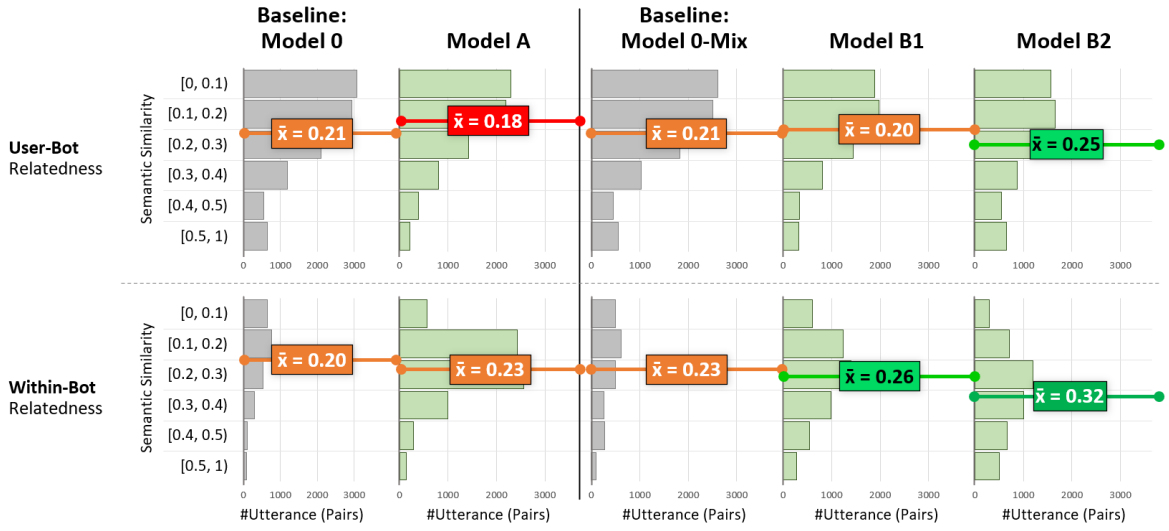


Figure 4.13: User-Bot Relatedness and Within-Bot Relatedness

As supporting evidence, the results of the automized assessment concur with those of

the human trial and reveal additional insights. In this step, two metrics of semantic similarity are assessed in Figure 4.13. It shows the *User-Bot Relatedness*, the similarity between the user prompt and the first sentence of the bot utterance, which can be interpreted as gauging the immediate relatedness of the bot's response to the user prompt. The performance differential for this metric is less stark than the target-guiding metrics assessed earlier, with two exceptions that are statistically significant. First, Model **A** underperforms its baseline by a small margin, despite the fact that its reward signal incentivized it to seek high semantic similarity between the user prompt and the first bot sentence. Second, Model **B2** again displays statistically significant higher relatedness compared to all other models, despite the fact that its reward function didn't include a signal for conversational smoothness. The other metric shown in Figure 4.13 is the *Within-Bot Relatedness*, the average semantic similarity between the sentences of the same utterance. Here, a clearer increase in performance between the trained models and their baselines can be observed, to the tune of +0.03 to 0.12 points of semantic similarity. This counts especially for Model **B2**, which outperforms its baseline by a marked ˜40%, increasing semantic similarity by 0.09 from 0.23 in the baseline Mixed Generation-Retrieval Model **0-Mix** to 0.32.

	Model **0**	Model **A**	Model **0-Mix**	Model **B1**	Model **B2**
Model **0**					
Model **A**	**0.00**				
Model **0-Mix**	0.75	**0.00**			
Model **B1**	0.01	**0.00**	0.04		
Model **B2**	**0.00**	**0.00**	**0.00**	**0.00**	

Table 4.2: p-Value Results of Two-Sided t-Test on *User-Bot Relatedness* (Automized Assessment)

	Model **0**	Model **A**	Model **0-Mix**	Model **B1**	Model **B2**
Model **0**					
Model **A**	**0.00**				
Model **0-Mix**	**0.00**	0.92			
Model **B1**	**0.00**	**0.00**	**0.00**		
Model **B2**	**0.00**	**0.00**	**0.00**	**0.00**	

Table 4.3: p-Value Results of Two-Sided t-Test on *Within-Bot Relatedness* (Automized Assessment)

Quality of English

To assess the models' quality of English, respondents are asked to agree on the grammatical correctness of the chatbots' responses. Figure 4.14 shows that the Mixed Generation-Retrieval Models **B1** and **B2** perform largely comparable to the baseline. Model **A**, which was trained to achieve the target using only Generation, displays significantly lower grammatical correctness, with ˜60% *disagreeing* or *strongly disagreeing* with the statement.

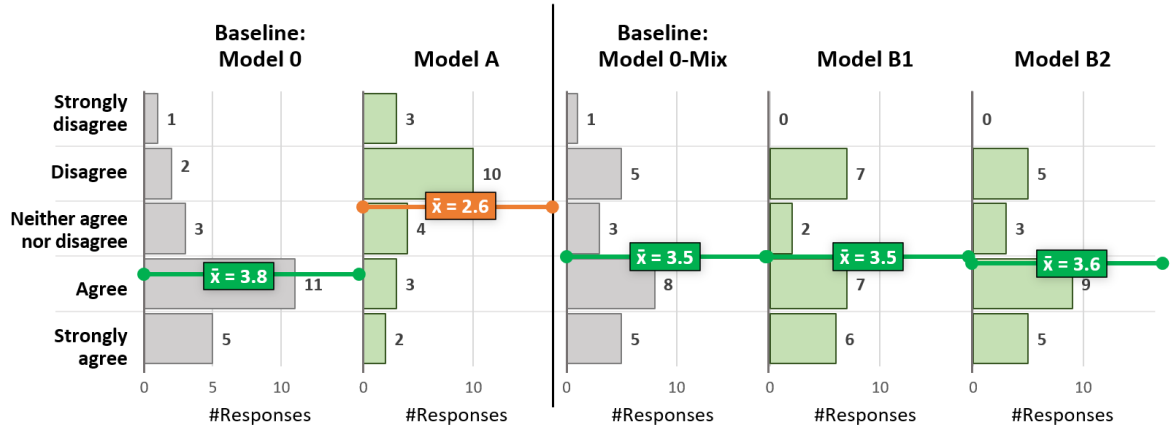


Figure 4.14: Participants' agreement with statement: '**The Chatbot's Responses were grammatically correct**'

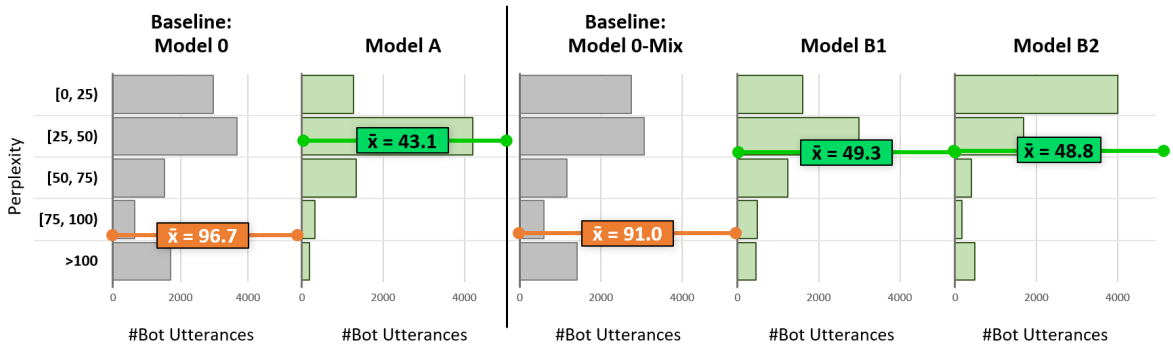


Figure 4.15: Perplexity of Bot Utterances (Automized Assessment)

The analysis of the perplexity of the bot utterances to assess a model's quality of English does not capture Model **A**'s lack of performance as described by the participants of the human trial. Instead, Figure 4.15 shows a strong decrease in perplexity for all models that were trained with Reinforcement Learning. Models **A**, **B1** and **B2** fall into a perplexity band of 43.1 - 49.3, each approximately with a 45-55% decrease over their baselines. These changes are highly statistically significant.

	Model **0**	Model **A**	Model **0-Mix**	Model **B1**	Model **B2**
Model **0**					
Model **A**	**0.00**				
Model **0-Mix**	0.56	**0.00**			
Model **B1**	**0.00**	**0.00**	**0.00**		
Model **B2**	**0.00**	0.02	**0.00**	0.85	

Table 4.4: p-Value Results of Two-Sided t-Test on Perplexity of Bot Utterance (Automized Assessment)

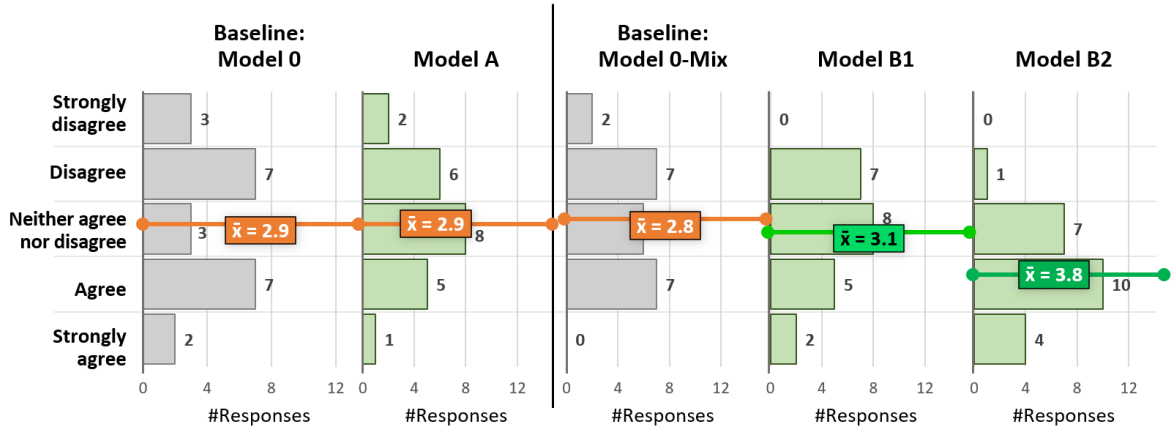


Figure 4.16: Participants' agreement with statement: '**The Chatbot's Responses were empathetic**'

Empathy

As a final note on the positive side-effects of the PPO training, the extent to which the models displayed empathy is examined in Figure 4.16. Similarly to the increase in relatedness and engagement described above, the strong increase in empathy appears unique to Model **B2**, with Models **A** and **B1** performing within a narrow 0.2 points of the Baseline Model **0**.

Summary of Comparative Conversational Smoothness

Comparing the models' performance on conversational smoothness demonstrates more clearly that the PPO approach followed with **B2** yields superior dialogue quality, level of English, and empathy compared to other approaches. No general performance increase can be identified stemming from Mixed Generation-Retrieval Models compared to Generation-only approaches. Especially the analysis of grammatical correctness highlights the performance shortcomings of Model **A**.

4.3 Discussion

As outlined in Chapter 2, the field of previous research applying PPO to conversational agents is scarce, and there exists no precedent for using PPO to optimize for target-guiding. The results of the research lend strong credibility to the proposed approach and prove that Proximal Policy Optimization is a tool to successfully implement target-guiding in open-domain dialogue systems with conversational goals. Section 4.1 demonstrated a large increase in performance especially when pairing the Reinforcement Learning training with a Mixed Generation-Retrieval strategy for utterance production. The best performing Model **B2** also demonstrates strong increases in dialogue quality and empathy relative to the baseline.

Using PPO instead of vanilla Reinforcement Learning for the optimization task provided more control over the balance of exploration and exploitation during training,

which was a key contributor to the quality of the models. Without controlling for proximity to the baseline model, which PPO does, trained models are expected to have much higher KL-Divergence, resulting in utterances with far lower quality of English. From the experiments and the fully developed models, one can conclude more generally that PPO especially in conjunction with a dedicated KL-penalty in the reward function, is a good tool for inducing behavior while maintaining language quality in dialogue systems.

The comparison of the trained models in Section 4.2 shows that when optimizing with PPO, the choice of reward signals and Language Model design matters. Specifically, it is observed that regarding rates of Task Success, Mixed Generation-Retrieval Models trained with PPO, i.e. Models **B1** and **B2**, perform best. When considering the models' conversational smoothness, optimizing with a reward signal focused solely on the target, i.e. Model **B2**, performs best. Two broader conclusions regarding PPO training for target-guiding are distilled, which also explain the differences in performance between the trained models.

Achieving Task Success: *Broad* Target with Language Model Adaptations

The target of the conversation, in this case having the user utter their emotional state, can be represented in myriad ways in the reward function used for PPO. A key trade-off to note here is how broadly the target is represented to the optimizer, as there are two conflicting objectives: First, in order to reliably reach the objective, this might require a very specific utterance to be produced, which would lead to narrowly defining the target as a *full sentence*, e.g. 'How are you feeling?'. Second, the flexibility when responding to the user should also be maintained, which is better attainable when the target is defined more broadly, e.g. producing an utterance around the *topic* of 'feelings'. Solely training the model this broadly, however, will likely not reliably reach the target of the user revealing their emotions, which is why modifications to the Language Model are required.

In Section 4.2, the implications of this trade-off are illustrated well by the differences in performance between Model **A** and the Mixed Generation-Retrieval Models **B1** and **B2**. Whereas the former contains a narrow target representation, the latter are optimized for a broader conversational topic and contain the mentioned Language Model modifications, where the question 'How are you feeling?' is appended to utterances that are sufficiently semantically related. It can be observed that the trained models with board target representations and Language Model modifications are more effective at target-guiding whilst offering increased conversational quality.

An additional explanation for this can be drawn from the feedback the PPO optimizer receives from the reward signal. Specifically, broad target representations are likely to generate higher proximity scores more quickly than narrow representations would. The reason for this is that a model's action a during an optimization step is more likely to contain a broad conversational topic that is close to the target topic, rather than containing an exact match with a complete target sentence. Hence, it will be easier for the PPO optimizer to detect high-reward behavior sooner and

thereby approximate its value function more accurately. It will be in a position to *exploit* this behavior much earlier, instead of being forced to *explore* a larger set of actions first. Excessively explorative behavior can have adverse consequences, such as losses in grammatical correctness as evidenced by Model **A** in Figure 4.14.

In fact, it is plausible that the bad grammar of Model **A** also negatively impacted its Task Success Rate in the human trial, where it showed no significant increase over its baseline. Its bad sentence structure is likely to have confused participants and deterred them from revealing their emotional state, something that did not occur during the automized trial where the user simulator tried to infer meaning from context.

As a final note, the use of semantic similarity, specifically the implementation by Cer et al. [34], appears to have worked well to produce a dense reward signal that reflects an utterance's proximity to the target. Its use both to reflect proximity between sentences and keywords proves its flexibility.

Achieving Conversational Smoothness: Focus on *Target* as Guiding Topic

For Mixed Generation-Retrieval models, different combinations of signals were trialed in the reward function. Whereas Models **B1** and **B2** both included a signal for target proximity and controlling the length of the utterance, only **B1** included an additional signal to optimize its utterances' conversational smoothness. In Section 4.2, it was shown that Model **B1** is strongly outperformed by Model **B2**, especially regarding the relatedness and the level of engagement of the dialogue.

Optimizing primarily for proximity to the target yields a better conversation quality by uninhibitedly giving the Language Model **A** a clear guiding conversational topic to focus on, without distracting the optimizer with a noisy auxiliary reward signal. As the baseline model had previously only been fine-tuned with general-purpose dialogue where topics may change quickly, the trained Model **B2** through its primary conversational topic is thereby able to outperform it both in terms of the conversation's relatedness and level of engagement. The average semantic similarity between sentences of the same bot utterance in Figure 4.13 provides evidence for this conclusion, where Model **B2** outperforms other models by a wide margin. Here, the guiding conversational topic of 'feelings' becomes evident by causing the higher semantic similarity between the sentences of the utterance. Other models do not demonstrate the same level of *Within-Bot Relatedness*, since they do not have a guiding conversation topic and the content of the conversation changes more quickly. Essentially, Model **B2**'s increased conversational smoothness is expected to be most predominant when the current conversation is close to or approaching **B2**'s guiding conversational topic. If participants of the trial had behaved unnaturally stubborn and had not followed Model **B2**'s inclination to move the conversation in the direction of 'feelings', **B2**'s response-relatedness would have likely been comparable with the baseline's relatedness.

Finally, on the combination of reward signals, controlling for side effects during optimization will still be necessary next to the reward signal for target-proximity.

As explained in Section 3.3.1, a small penalty was required to control the length of the generated utterances, setting the desired number of sentences per utterance to 2. Since this penalty did not carry a large weight in the reward function, it did not distract the optimizer from seeking proximity to the target. It likely even contributed to the increased level of engagement of the dialogue of trained models compared to the baseline, as the baseline's utterances were sometimes short and hence did not stimulate the conversation.

4.4 Limitations and Future Work

This section reflects on the limitations of the chosen implementation and proposes future work for the continued development of the SAT-Chatbot.

Limitations of the PPO Research

Whereas the assessment of this research was comprehensive in capturing many different metrics for task success and conversational smoothness, future implementations could improve the accuracy of how these results were captured. Specifically, as described in Section 3.5, task success was measured by analyzing user responses for the presence of keywords (e.g. 'happy', 'sad') that are connected with their emotional state. Whereas this approach was found to detect the target more reliably than using existing emotion classifiers, solely scanning for keywords will not detect emotions with perfect accuracy, as it does not account for, e.g., a user's implied intent or spelling mistakes. The results of the human trial and manually vetting chat transcripts lends confidence to the target-guiding effect of the models, however using e.g. improved dedicated classifiers to detect information about a user's emotion would improve the accuracy of the Task Success Rate. Future implementations of PPO for target-guiding should also validate the target-guiding performance using a more stable baseline. Since the current SAT-Chatbot often has poor dialogue quality, the final conversational smoothness of the optimized model is still constrained by its baseline, despite the significant relative performance increases of this research. Such validation of the results could involve applying the PPO approach to state-of-the-art chit-chat agents such as Microsoft's *DialoGPT* [54].

The PPO approach holds additional potential for target-guiding that could be unlocked by making more fundamental changes to the RL-components introduced in Section 3.3.2. Specifically, an action a was defined as a *single-turn* response to a user utterance, which was rewarded for increasingly approaching the target, implicitly resulting in a *multi-turn* execution of dialogue strategy. For further refinement of the PPO approach, one could explicitly train for *multi-turn* responses already during PPO optimization, enabling the Language Model to better plan topic transitions across a conversation history.

Continued Development of the SAT-Chatbot

Outside the realm of this research, the SAT-Chatbot should be further refined to eventually be used in a clinical setting. Concrete recommendations are made for the improvement of the conversational quality and increasing control over dialogue flow, which can be achieved using Symbolic AI.

Improving Dialogue Quality As mentioned above, whereas this research demonstrated strong *relative* increases in dialogue quality, the conversational smoothness of Model **B2** is still constrained in absolute terms. Hence, to yield an overall improved model, higher conversational smoothness could be achieved by two means: First, the Transformer architecture could be upgraded from GPT-2 to GPT-3, which was recently made publicly available and outperforms its predecessor in myriad NLP tasks [58]. Second, the fine-tuning of the pre-trained Transformers could be performed on dialogue corpora that avoid frequent topic changes more consistently, such as PersonaChat [59], Wizard of Wikipedia [60], and BlendedSkillTalk [61]. Once dialogue quality is improved, the flexibility of the PPO approach proposed in this research will make it easy to adjust the implementation to introduce target-guiding to an optimized baseline of the SAT-Chatbot.

Exerting more control over Dialogue Flow using Symbolic AI The current SAT-Chatbot generates dialogue using Machine Learning, which produces flexible responses but forfeits control over the produced utterances. As an alternative, Symbolic AI could be used to exert more control over the dialogue, which is highly attractive given the sensitive application of this bot in psychotherapy. It is based on two aspects: First, it embeds inputs into symbolic representations that can be interpreted by humans. Second, its behavior is based on applying explicitly defined rules to its inputs using Logic, which is defined by humans [62]. Hence, whereas behavior is *learned* in Machine Learning, it is *prescribed* and *explainable* in Symbolic AI. A possible application of this is illustrated by Wu et al's approach to target-guiding [43] which was already briefly mentioned in Section 2.5. Specifically, they designed a Language Model that combines Symbolic AI for defining the dialogue strategy, with Machine Learning for generating the final utterance. As a symbolic representation of the dialogue strategy, they create a knowledge network, i.e. a curated network of conversational topics that are connected through knowledge relations[1]. When receiving a user prompt, the model draws a knowledge graph between the user's input and the conversational target, which results in a path of topics used to generate utterances over a number of turns, eventually arriving at the target. As a key benefit, the backbone knowledge network and the rules by which to form paths through this network are fully explainable and controllable [43]. For the continued development of the SAT-Chatbot in the Emotion Recognition Stage, such prescribed dialogue paths could be used to further increase the efficiency of the target-guiding and to prevent the chatbot from discussing sensitive topics. Symbolic AI could also be applied to represent more sophisticated rules in the Protocol Recommendation

[1] E.g. Relation between 'Harry Potter and the Philosopher's Stone' & 'Chris Columbus': 'directed-by'

Stage of the SAT-Chatbot. Currently, in this stage, the bot asks clarifying questions and gives protocol recommendations mostly by accepting simple 'Yes/No' inputs. When implementing more sophisticated features, their implementation rules could be represented effectively using Logic and Symbolic AI.

Miscellaneous further Ideas on future SAT-Chatbot Development Finally, future projects to continue the work on the SAT-Chatbot more generally include introducing dialogue state tracking for the same patients *over multiple sessions*, integrating an *audio interface* to the chatbot that also uses voice analysis to improve emotion recognition, and extending the scope of the chatbot to not just recommend protocols but also *engage with the inner child*.

Chapter 5

Conclusion

This work presents a novel approach of using Reinforcement Learning with Proximal Policy Optimization to train the SAT-Chatbot to reliably guide the conversation towards its designated goal, capturing the patient's emotional state. The optimized chatbot is highly effective in achieving its target, reaching it within 20 turns in up to 86% of interactions with users. It also reaches the target more efficiently, in up to 30% fewer turns compared to the baseline. As a positive side-effect, this further significantly improves the often poor conversational quality of the baseline, with respondents of the human trial on average agreeing that their interaction with the optimized model had high relatedness and engagement. Finally, the approach also improved the level of empathy of the SAT-Chatbot.

Our research proves that PPO is an effective tool for achieving such well-performing target-guiding. The implications of alternative PPO training configurations are analyzed more broadly in terms of reward signals and Language Model design over a comprehensive and differentiated assessment of the developed models. This is used to derive general conclusions for applying Proximal Policy Optimization to conversational agents, which is a highly scarce area of research:

- **Achieving Task Success**: Optimize for a *broad* target representation and make modifications to the Language Model design accordingly
- **Achieving Conversational Smoothness**: Focus reward signal on the *target* to give the dialogue a guiding topic, which improves conversational quality

In the context of the SAT-Chatbot, this project has markedly improved the Emotion Recognition Stage of the bot-patient interaction and has brought the bot one step closer to a clinical trial and its eventual use in providing digital psychotherapy. More generally in the context of NLP, this research can help inform future applications of PPO, which could be used to train dialogue systems to achieve even more complex targets.

Appendix A

Overview of SAT-Protocols

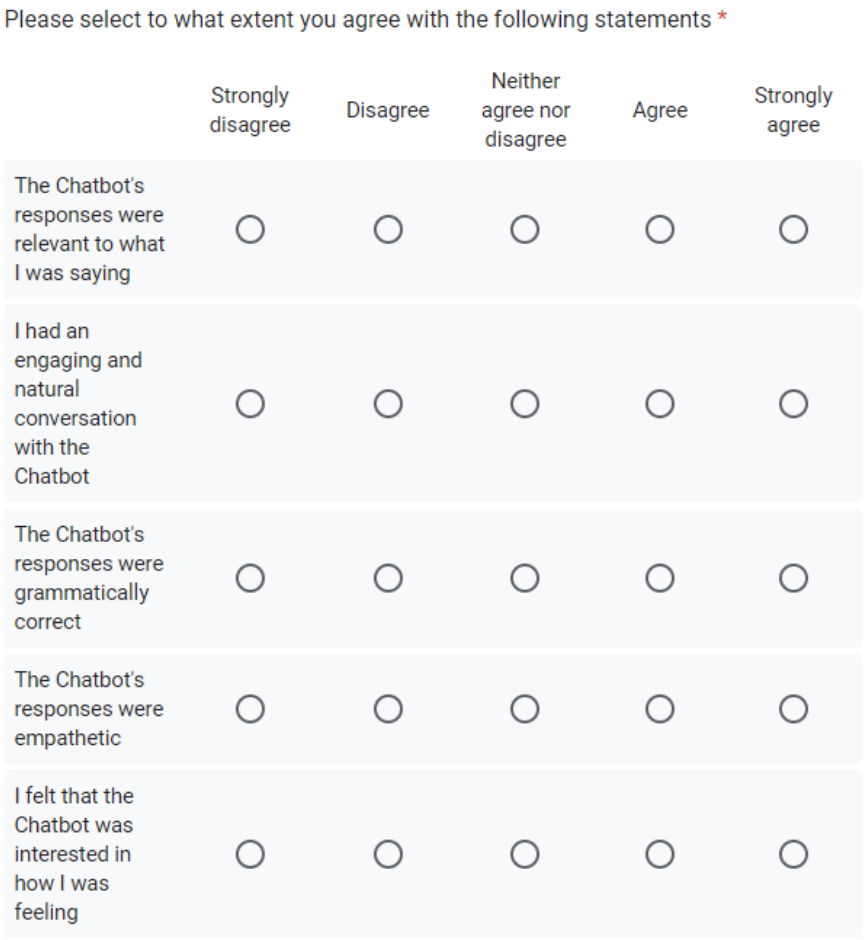
1. Recalling significant early memories
2. Becoming intimate with your Child
3. Singing a song of affection
4. Expressing love and care for the Child
5. Pledging to care and support our Child
6. Restoring our emotional world after our pledge
7. Maintaining a loving relationship with your Child, Creating Zest for life
8. Enjoying nature
9. Overcoming your current negative emotions
10. Overcoming past pain
11. Muscle relaxation and playful face
12. Laughing on your own
13. Laughing with your childhood self
14. Creating your own brand of laughter
15. Learning to change your perspective
16. Learning to be playful about your past pains
17. Identifying our personal resentments and acting them out
18. Planning more constructive actions
19. Updating our beliefs to enhance creativity
20. Practicing affirmations

Taken from *Self-attachment Technique (SAT): Detailed Exercises - Version 6* [63].

Appendix B

Human Trial Questionnaire

In the human trial, participants were asked to chat with five different chatbots, answering its first question 'What did you do today?' the same way every time. After each interaction, they answered the same five questions:



Please select to what extent you agree with the following statements *

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The Chatbot's responses were relevant to what I was saying	○	○	○	○	○
I had an engaging and natural conversation with the Chatbot	○	○	○	○	○
The Chatbot's responses were grammatically correct	○	○	○	○	○
The Chatbot's responses were empathetic	○	○	○	○	○
I felt that the Chatbot was interested in how I was feeling	○	○	○	○	○

Figure B.1: Questionnaire after Chat Interaction in Human Trial

Bibliography

[1] James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. A systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 11;392:1789-858. Available from: https://doi.org/10.1016/S0140-6736(18)32279-7. pages 1

[2] Rudd BN, Beidas RS. Digital Mental Health: The Answer to the Global Mental Health Crisis? JMIR Mental Health. 2020 6;7:e18472. Available from: https://mental.jmir.org/2020/6/e18472. pages 1

[3] Rimmer A. Mental health: Staff shortages are causing distressingly long waits for treatment, college warns. BMJ. 2021 10:n2439. Available from: https://www.bmj.com/content/375/bmj.n2439. pages 1

[4] McDonald A, Eccles JA, Fallahkhair S, Critchley HD. Online psychotherapy: trailblazing digital healthcare. BJPsych Bulletin. 2020 4;44:60-6. Available from: http://dx.doi.org/10.1192/bjb.2019.66. pages 1, 3

[5] Edalat A. Self-attachment: A holistic approach to computational psychiatry. Computational neurology and psychiatry. 2017:273-314. Available from: http://dx.doi.org/10.1007/978-3-319-49959-8_10. pages 1, 4

[6] Alazraki L, Ghachem A, Polydorou N, Khosmood F, Edalat A. An Empathetic AI Coach for Self-Attachment Therapy. 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI). 2021 12:78-87. Available from: https://doi.org/10.1109/CogMI52975.2021.00019. pages 1, 4, 23

[7] Yan H. A virtual psychotherapist that can understand human language. 2022. MEng Thesis. pages 1, 4, 5, 11, 23

[8] NHS. The NHS Long Term Plan; 2019. Available from: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf. pages 3

[9] Weightman M. Digital psychotherapy as an effective and timely treatment option for depression and anxiety disorders: Implications for rural and remote practice. Journal of International Medical Research. 2020 6;48. Available from: https://doi.org/10.1177/0300060520928686. pages 3

[10] Hawa S, Akella S, Kaushik S, Joshi V, Kalbande D. Analysis of Therapy Transcripts using Natural Language Processing. International Journal of Engi-

neering and Advanced Technology. 2020 8;9:489-94. Available from: https://doi.org/10.35940/ijeat.F1598.089620. pages 3

[11] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966 1;9:36-45. Available from: https://doi.org/10.1145/365153.365168. pages 3

[12] Xu B, Zhuang Z. Survey on psychotherapy chatbots. Concurrency and Computation: Practice and Experience. 2022 3;34. Available from: https://doi.org/10.1002/cpe.6170. pages 3, 27

[13] Yin J, Chen Z, Zhou K, Yu C. A Deep Learning Based Chatbot for Campus Psychological Therapy. CoRR. 2019;abs/1910.06707. Available from: https://doi.org/10.48550/arXiv.1910.06707. pages 3

[14] Huang J, Li Q, Xue Y, Cheng T, Xu S, Jia J, et al. TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress. International Conference on Health Information Science. 2015 8:133-45. Available from: https://doi.org/10.1007/978-3-319-19156-0_14. pages 3

[15] Mikulincer M, Shaver PR. An attachment perspective on psychopathology. World Psychiatry. 2012 2;11:11-5. Available from: https://doi.org/10.1016%2Fj.wpsyc.2012.01.003. pages 4

[16] Jurafsky D, Martin JH. Speech and Language Processing. 3rd ed. Prentice Hall; 2022. pages 5, 6

[17] Manning C. Language Models and Recurrent Neural Networks - Lecture Notes. CS224n: Natural Language Processing with Deep Learning - Lecture Notes. 2022. Available from: https://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture05-rnnlm.pdf. pages 5

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30. Available from: https://doi.org/10.48550/arXiv.1706.03762. pages 5

[19] Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. OpenAI Blog. 2018. Available from: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. pages 5, 7

[20] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 6:4171-86. Available from: http://dx.doi.org/10.18653/v1/N19-1423. pages 5

[21] Li Y. Reinforcement learning applications. arXiv preprint arXiv:190806973.

2019. Available from: https://doi.org/10.48550/arXiv.1908.06973. pages 5, 7

[22] Li Y. Deep Reinforcement Learning. CoRR. 2018;abs/1810.06339. Available from: https://doi.org/10.48550/arXiv.1810.06339. pages 5, 7

[23] Young T, Xing F, Pandelea V, Ni J, Cambria E. Fusing Task-Oriented and Open-Domain Dialogues in Conversational Agents. Proceedings of the AAAI Conference on Artificial Intelligence. 2022 6;36:11622-9. Available from: https://doi.org/10.48550/arXiv.2109.04137. pages 5

[24] Adewumi T, Liwicki F, Liwicki M. State-of-the-Art in Open-Domain Conversational AI: A Survey. Information. 2022 6;13:298. Available from: http://dx.doi.org/10.20944/preprints202205.0016.v1. pages 6

[25] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002 7:311-8. Available from: http://dx.doi.org/10.3115/1073083.1073135. pages 6

[26] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. 2004 7:74-81. Available from: https://aclanthology.org/W04-1013.pdf. pages 6

[27] Gupta P, Jhamtani H, Bigham JP. Target-Guided Dialogue Response Generation Using Commonsense and Data Augmentation. ArXiv. 2022;abs/2205.09314. Available from: https://doi.org/10.48550/arXiv.2205.09314. pages 6, 8, 9, 10, 18, 19, 20

[28] Liu CW, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 11:2122-32. Available from: https://doi.org/10.48550/arXiv.1603.08023. pages 6, 7, 16, 24

[29] Deriu J, Rodrigo A, Otegi A, Echegoyen G, Rosset S, Agirre E, et al. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review. 2021 1;54:755-810. Available from: https://doi.org/10.1007%2Fs10462-020-09866-x. pages 6

[30] See A, Roller S, Kiela D, Weston J. What makes a good conversation? How controllable attributes affect human judgments. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 6:1702-23. Available from: http://dx.doi.org/10.18653/v1/N19-1170. pages 6

[31] Chandrasekaran D, Mago V. Evolution of Semantic Similarity—A Survey. ACM

Computing Surveys. 2022 3;54:1-37. Available from: https://doi.org/10.1145/3440755. pages 6

[32] Tang J, Zhao T, Xiong C, Liang X, Xing EP, Hu Z. Target-Guided Open-Domain Conversation. CoRR. 2019;abs/1905.11553. Available from: https://doi.org/10.48550/arXiv.1905.11553. pages 6, 8, 9, 10, 14, 19

[33] Qin J, Ye Z, Tang J, Liang X. Dynamic Knowledge Routing Network For Target-Guided Open-Domain Conversation. CoRR. 2020;abs/2002.01196. Available from: https://doi.org/10.1609/aaai.v34i05.6390. pages 6, 9, 10

[34] Cer D, Yang Y, yi Kong S, Hua N, Limtiaco N, John RS, et al. Universal Sentence Encoder. CoRR. 2018;abs/1803.11175. Available from: https://doi.org/10.48550/arXiv.1803.11175. pages 6, 13, 14, 15, 19, 41

[35] Zhang Z, Takanobu R, Huang M, Zhu X. Recent Advances and Challenges in Task-oriented Dialog System. CoRR. 2020;abs/2003.07490. Available from: https://doi.org/10.48550/arXiv.2003.07490. pages 6

[36] Lee S, Zhu Q, Takanobu R, Zhang Z, Zhang Y, Li X, et al. ConvLab: Multi-Domain End-to-End Dialog System Platform. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019 7:64-9. Available from: http://dx.doi.org/10.18653/v1/P19-3011. pages 6

[37] Sutton R, Barto A. Reinforcement Learning: An Introduction. 2nd ed. The MIT Press; 2014. pages 6, 7

[38] Ciaming X. Reinforcement Learning for NLP. CS11-747 Neural Networks for NLP - Lecture Notes. 2018. Available from: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/lectures/lecture16-guest.pdf. pages 7

[39] Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, et al. Decision Transformer: Reinforcement Learning via Sequence Modeling. Advances in Neural Information Processing Systems. 2021;34:15084-97. Available from: https://doi.org/10.48550/arXiv.2106.01345. pages 7

[40] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms. CoRR. 2017;abs/1707.06347. Available from: https://doi.org/10.48550/arXiv.1707.06347. pages 8

[41] Ziegler D, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al. Fine-Tuning Language Models from Human Preferences. CoRR. 2019;abs/1909.08593. Available from: http://arxiv.org/abs/1909.08593. pages 8, 13, 14, 18, 25

[42] Faal F, Yu JY, Schmitt K. Transformer Decoder Based Reinforcement Learning Approach for Conversational Response Generation. 2020 International Joint Conference on Neural Networks (IJCNN). 2020:1-8. Available from: http://dx.doi.org/10.1109/IJCNN48605.2020.9207289. pages 8, 13, 19

[43] Wu W, Guo Z, Zhou X, Wu H, Zhang X, Lian R, et al. Proactive Human-Machine Conversation with Explicit Conversation Goals. CoRR. 2019;abs/1906.05572. Available from: https://doi.org/10.48550/arXiv.1906.05572. pages 9, 43

[44] Zhou L, Small K, Rokhlenko O, Elkan C. End-to-End Offline Goal-Oriented Dialog Policy Learning via Policy Gradient. CoRR. 2017;abs/1712.02838. Available from: https://doi.org/10.48550/arXiv.1712.02838. pages 9, 10

[45] Peng B, Li X, Li L, Gao J, Celikyilmaz A, Lee S, et al. Composite Task-Completion Dialogue System via Hierarchical Deep Reinforcement Learning. CoRR. 2017;abs/1704.03084. Available from: https://doi.org/10.48550/arXiv.1704.03084. pages 9, 10

[46] Lipton ZC, Gao J, Li L, Li X, Ahmed F, Deng L. Efficient Exploration for Dialog Policy Learning with Deep BBQ Networks & Replay Buffer Spiking. CoRR. 2016;abs/1608.05081. Available from: https://doi.org/10.48550/arXiv.1608.05081. pages 9

[47] Papangelis A, Wang YC, Molino P, Tür G. Collaborative Multi-Agent Dialogue Model Training Via Reinforcement Learning. CoRR. 2019;abs/1907.05507. Available from: http://dx.doi.org/10.18653/v1/W19-5912. pages 9

[48] Liu B, Tür G, Hakkani-Tür D, Shah P, Heck LP. End-to-End Optimization of Task-Oriented Dialogue Model with Deep Reinforcement Learning. CoRR. 2017;abs/1711.10712. Available from: https://doi.org/10.48550/arXiv.1711.10712. pages 9

[49] Dhingra B, Li L, Li X, Gao J, Chen YN, Ahmed F, et al. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. CoRR. 2016;abs/1609.00777. Available from: https://doi.org/10.48550/arXiv.1609.00777. pages 9

[50] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1:9. Available from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. pages 11

[51] Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. CoRR. 2017;abs/1710.03957. Available from: https://doi.org/10.48550/arXiv.1710.03957. pages 12

[52] Ritter A, Cherry C, Dolan WB. Data-Driven Response Generation in Social Media. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011 7:583-93. Available from: https://aclanthology.org/D11-1054. pages 12

[53] Wang H, Lu Z, Li H, Chen E. A dataset for research on short-text conversation. EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2013 8:935-45. Available from: https://aclanthology.org/D13-1096. pages 12

[54] Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, et al. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. CoRR. 2019;abs/1911.00536. Available from: http://dx.doi.org/10.18653/v1/2020.acl-demos.30. pages 14, 23, 42

[55] Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword Extraction from Single Documents using Multiple Local Features. Information Sciences. 2020 8;509:257-89. Available from: https://doi.org/10.1016/j.ins.2019.09.013. pages 19

[56] Britannica E. Words for Emotions: Vocabulary Words; 2022. Available from: https://www.britannica.com/dictionary/eb/3000-words/topic/emotions-vocabulary-english/1. pages 23

[57] London IC. Legal, Social, Ethical and Professional Requirements: Ethics Checklist; 2020. Available from: https://wiki.imperial.ac.uk/display/docteaching/Legal%2C+Social%2C+Ethical+and+Professional+Requirements. pages 26

[58] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. CoRR. 2020;abs/2005.14165. Available from: https://doi.org/10.48550/arXiv.2005.14165. pages 43

[59] Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing Dialogue Agents: I have a dog, do you have pets too? CoRR. 2018;abs/1801.07243. Available from: https://doi.org/10.48550/arXiv.1801.07243. pages 43

[60] Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J. Wizard of Wikipedia: Knowledge-Powered Conversational agents. CoRR. 2018;abs/1811.01241. Available from: https://doi.org/10.48550/arXiv.1811.01241. pages 43

[61] Smith EM, Williamson M, Shuster K, Weston J, Boureau YL. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. CoRR. 2020;abs/2004.08449. Available from: https://doi.org/10.48550/arXiv.2004.08449. pages 43

[62] Lima GF, Costa R, Moreno MF. An Introduction to Symbolic Artificial Intelligence Applied to Multimedia. CoRR. 2019;abs/1911.09606. Available from: https://doi.org/10.48550/arXiv.1911.09606. pages 43

[63] Group AHD. Self-attachment Technique (SAT): Detailed Exercises - Version 6; 2022. pages 46