electronics

MDPI

Article

AI to Train AI: Using ChatGPT to Improve the Accuracy of a Therapeutic Dialogue System

Karolina Gabor-Siatkowska [1,*], Marcin Sowański [1], Rafał Rzatkiewicz [1], Izabela Stefaniak [2], Marek Kozłowski [1] and Artur Janicki [1,*]

[1] Faculty of Electronics and Information Technology, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland; marcin.sowanski.dokt@pw.edu.pl (M.S.); rafal.rzatkiewicz.stud@pw.edu.pl (R.R.); marek.kozlowski@pw.edu.pl (M.K.)

[2] Faculty of Medicine, Lazarski University, ul. Świeradowska 43, 02-662 Warsaw, Poland; istefaniak.terapia@gmail.com

* Correspondence: karolina.gabor-siatkowska.dokt@pw.edu.pl (K.G.-S.); artur.janicki@pw.edu.pl (A.J.)

Abstract: In this work, we present the use of one artificial intelligence (AI) application (ChatGPT) to train another AI-based application. As the latter one, we show a dialogue system named Terabot, which was used in the therapy of psychiatric patients. Our study was motivated by the fact that for such a domain-specific system, it was difficult to acquire large real-life data samples to increase the training database: this would require recruiting more patients, which is both time-consuming and costly. To address this gap, we have employed a neural large language model: ChatGPT version 3.5, to generate data solely for training our dialogue system. During initial experiments, we identified intents that were most often misrecognized. Next, we fed ChatGPT with a series of prompts, which triggered the language model to generate numerous additional training entries, e.g., alternatives to the phrases that had been collected during initial experiments with healthy users. This way, we have enlarged the training dataset by 112%. In our case study, for testing, we used 2802 speech recordings originating from 32 psychiatric patients. As an evaluation metric, we used the accuracy of intent recognition. The speech samples were converted into text using automatic speech recognition (ASR). The analysis showed that the patients' speech challenged the ASR module significantly, resulting in deteriorated speech recognition and, consequently, low accuracy of intent recognition. However, thanks to the augmentation of the training data with ChatGPT-generated data, the intent recognition accuracy increased by 13% relatively, reaching 86% in total. We also emulated the case of an error-free ASR and showed the impact of ASR misrecognitions on the intent recognition accuracy. Our study showcased the potential of using generative language models to develop other AI-based tools, such as dialogue systems.

Keywords: spoken dialogue system; speech recognition; ChatGPT; data augmentation; computer-aided therapy; cognitive-behavioral therapy

Citation: Gabor-Siatkowska, K.; Sowański, M.; Rzatkiewicz, R.; Stefaniak, I.; Kozłowski, M.; Janicki, A. AI to Train AI: Using ChatGPT to Improve the Accuracy of a Therapeutic Dialogue System. *Electronics* **2023**, *12*, 4694. https://doi.org/10.3390/electronics12224694

Academic Editors: Seifedine Kadry, Isah A. Lawal and Sahar Yassine

Received: 5 August 2023
Revised: 9 November 2023
Accepted: 16 November 2023
Published: 18 November 2023

1. Introduction

The growing number of people suffering from various mental disorders, such as depression, anxiety, phobias, or schizophrenia, is one of the greatest challenges of contemporary societies. It is unsurprising that new technologies, including artificial intelligence (AI), are being tried to help in this area. One example of the use of AI technologies, in this case, natural language processing (NLP), automatic speech recognition (ASR), and machine learning (ML), is a dialogue system helping in therapy for mental illnesses, called Terabot [1]. By interacting with a dialogue agent, psychiatric patients suffering from complex, overwhelming emotions such as anxiety, anger, shame, or frustration can learn to control them while receiving their usual treatment. During a conversation with Terabot, patients are helped to calm their aroused emotions and are encouraged to perform a relaxing exercise.

While initial experiments with Terabot were very promising [2], they also revealed several challenges caused, among other things, by the quality of input data. The input speech of psychiatric patients is often slurred (e.g., as a side effect of aggravating pharmacological treatment), so the ASR system has problems recognizing it correctly. Consequently, this often leads to incorrect recognition of the speaker's intent.

A remedy for this would be enlarging the size of the dataset used to train the intent recognition module of the dialogue system. Collecting real-life users' data (here: patients' utterances from their therapeutic sessions) is very inconvenient; it is time-consuming and costly. In addition, it raises ethical questions, as the patients would need to work with a dialogue system of inferior quality. Therefore, to fill this gap, in this work, we propose using ChatGPT to augment the training dataset, aiming to improve intent recognition of our dialogue system.

The idea of using AI tools to develop other AI tools has been researched for some time. In natural language understanding (NLU), data augmentation has been a widely used technique, as it is known that neural models work better when presented with many examples per class. Recently with the introduction of large-scale language models (LLMs), data augmentation for NLU has become more elaborate as more meaningful examples can be generated via prompt-based learning. In [3], the authors proposed LINGUIST, a method for generating annotated data for intent recognition and slot filling. The method uses 20B-parameter LLMs to generate new examples guided by structured input that defines the expected output. The method surpasses classic augmentation techniques of Back-Translation and Example Extrapolation by 1.9% relative on intent recognition recall and 2.5% relative on slot-filling F1 score.

PLACES [4] is another model that synthesizes full natural language conversations using the 30B-parameter model for that purpose. Synthetic conversations generated by the presented method are comparable in terms of quality and lexical diversity to the datasets created by human annotators. Prompt-based data augmentation was also used to generate emotional support conversation. AugESC [5] uses a 6B-parameter to increase the training corpus's size and cover new topics.

In psychiatry and psychology, new technologies have supported or supplemented the therapist's work in various ways. For example, some tools employ various forms of cognitive-behavioral therapy (CBT) or other techniques e.g., exposure in virtual reality settings [6]. Human-controlled dialogue systems were successfully trialed in therapy for auditory hallucinations [7,8]. In these studies, patients were exposed to on-screen avatars, which emulated patients' hallucinations and helped cope with them. It has also been shown that computer-based techniques , such as virtual and augmented reality, avatars, and brain–computer interfaces [9], can strongly support therapy for mental disorders.

Dialogue systems and chatbots have been the main focus of AI ever since the term was coined in 1956. Ten years later, Weizenbaum created a rule-based chatbot, ELIZA [10], which simulated a Rogerian psychotherapist. Until the late 1990s and early 2000s, chatbots were mostly rule-based. Then, in the mid-2000s, machine learning techniques were introduced, and in the mid-2010s, deep-learning methods. Both significantly improved the quality of chatbots.

Goal-oriented dialogue systems, also known as conversation systems, aim to fulfill a specific task by talking to the user. Typically, such a system consists of a few sub-systems: NLU with intent recognition and slot filling [11], dialogue state tracking [12], dialogue management [13], and language generation [14]. For example, a Woebot chatbot, installed on a mobile application, turned out to be helpful in therapy for depression [15]. The results of the interaction with the Woebot were very promising, as it turned out that after 2 weeks, it was already possible to see reduced depression in those who participated. Another example is ADELE, an artificial conversational care agent, which has been used in a social robot helping elderly people [16]. Its aim was to take care of the elderly by monitoring their health and well-being through social dialog. The use of dialogue systems in social robots is discussed, among other aspects, in [17]. Insight is provided into different perspectives

of interaction, including interaction in social space, interaction in groups, and interaction over time.

In the field of dialogue systems, the recognition of intentions and emotions is of crucial importance. A dialogue agent must accurately interpret the intentions and emotions of the speaker to provide a suitable response. Our team addressed this issue by performing sentiment and emotion recognition experiments on English and Polish texts in the context of a therapeutic chatbot [18]. First, an existing English-language corpus labeled with emotions was augmented with neutral text samples. Next, it was machine-translated into Polish. The resulting bilingual parallel corpus was then used for classification experiments. It was labeled with three sentiment polarity classes and nine emotion classes. The best results were obtained for the models based on Bidirectional Encoder Representations from Transformers (BERT).

Research on emotion recognition is still ongoing, as presented, e.g., in [19]. To capture the emotions of online communication, the researchers used short Twitter texts. They proposed a new semi-automatic method for multi-labeling short Internet texts (tweets) and built a multi-label corpus for training the algorithm. Another study [20] goes a step further. In their research, a multi-label *k* nearest neighbor (MLkNN) classifier was modified to allow iterative corrections for multi-label emotion classification by considering not only individual features within sentences, but also adjacent sentences and the entire tweet. The findings presented in their paper show that through this approach, the accuracy and speed of emotion classification in Twitter texts increase.

Recent development in conversational AI has been led almost exclusively by transformer-based neural language models specializing in dialogue generation. Since the initial release of the Transformer [21], or more specifically, the BERT model [22], improvements to dialogue generation have been made mainly through increasing the number of network parameters. This strategy has proven so effective that in just a couple of years, the set of problems in state-of-the-art models has shifted from problems with short, dull, and uncontrollable answers [23,24] to problems of how much understanding of meaning these models have [25,26].

LLMs themselves represent a significant advance in AI. From different perspectives, such as education, industry or medicine, researchers discover and show the potential benefits and challenges of different applications of LLMs [27]. Currently, the state-of-the-art conversational LLM model is ChatGPT [28], built on the GPT-4 model. It is capable of generating long and consistent responses on a variety of topics.

There has also been much discussion about using ChatGPT for medical applications. For example, in [29] the authors compared ChatGPT with existing tools for ophthalmic diagnosis, namely Isabel Pro differential diagnosis generator. After performing various analyses, their study concluded that conversational AI models like ChatGPT have potential value in diagnosing ophthalmic conditions. ChatGPT has also been reported as successful in fulfilling supportive tasks in medical care, e.g., generating valuable suggestions for clinical decision support logic [30].

Still, there are serious concerns about the trustworthiness of the generated responses and the problem of hallucination of the model [31]. In [32], researchers showed that even though the hallucinated text is false and nonsensical, it gives the user the impression that it is fluent and natural. It has the appearance of being based on a real context, although it is actually difficult to verify the existence of such contexts. Hallucinations in LLMs are very similar to psychological hallucinations: they both are difficult to distinguish from genuine perceptions. Hallucinations may result in spreading misinformation, exposing confidential information, and creating unrealistic expectations about LLM capabilities.

In this work, we propose improving the accuracy of intent recognition of our therapeutic dialogue system by employing ChatGPT to generate additional training data. In contrast to employing a large neural language model as an engine of the dialogue system (which could give rise to the risk of model hallucinations, i.e., generating random utterances), we will use ChatGPT solely to augment training data for the current, much

simpler, goal-oriented system. This way, we will improve the system yet retain control over its behavior.

The rest of our paper is structured as follows. In Section 2.1 we describe the therapeutic spoken dialogue system, while its ChatGPT-driven training is described in Section 2.2. Our experiments are outlined in Section 3. Their results are shown in Section 4 and discussed in Section 5. This paper concludes with a summary in Section 6.

2. Material and Methods

2.1. Terabot—Therapeutic Spoken Dialogue System

In this study, a spoken dialogue agent called Terabot was equipped with a voice interface that operates in the Polish language, as it is supposed to work with Polish-speaking patients. The system was designed for psychiatric therapy; therefore, it was not based on a neural, Transformer-based system given its lack of controllability [31,33] and the large amount of training data required. Instead, to minimize the risk of giving the patient an inappropriate answer, we decided to use a goal-directed system. Figure 1 shows a schematic diagram of Terabot. When a patient talks, the speech is converted into text by a text-to-speech system (ASR). This was carried out using the Google Web Speech API in the Polish language.

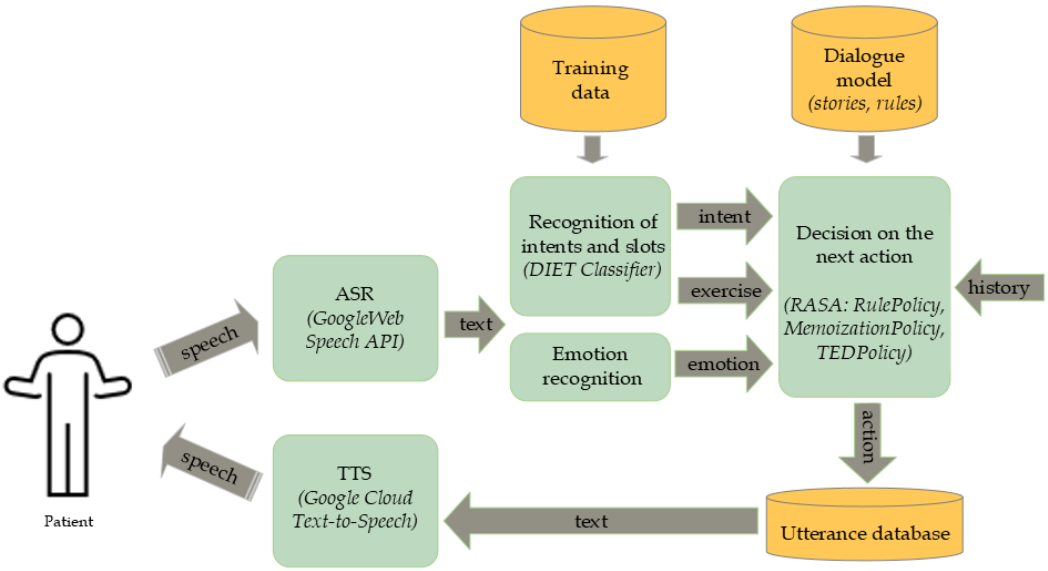

Figure 1. Block diagram of Terabot dialogue system.

Next, the text is analyzed using the Dual Intent and Entity Transformer (DIET) classifier [34]—it identifies the intents and slots in the patient's statements. Examples of intents recognized by the dialogue system are *Choose_exercise* for selecting an exercise or *Say_story* for a patient telling a story. A few separate intents are used for chatting or answering some basic questions, such as *Chitchat* and *Faq* (see Table 1 for a list of other upcoming intents). To implement the action decision pipeline and the DIET classifier, we used an open-source framework for NLU and dialogue management named RASA 3.1 (https://rasa.com/product/rasa-platform/, accessed on 4 August 2023) [35].

Simultaneously with intent and slot recognition, the ASR output is fed into the text-based emotion-recognition module [1,18]. This module is based on the BERT model [22] and was then fine-tuned to the emotion-classification task. The current emotional state of the patient is detected, and for this, the value of the *emotion* slot is set. A different slot is filled with the exercise type. In the next step, based on the combination, a decision is made about the next system action to be taken: memorization policy (i.e., based on stories kept in memory), a rule policy, and a Transformer Embedding Dialogue (TED) policy [36]. The TED policy takes into account the current state of the dialogue, such as the patient's intent,

slot values (including a recognized emotion), previous states of the dialogue, etc. If the next selected action is an utterance, it is selected from the utterance database prepared according to the psychiatrist's instructions. The Google Cloud Text-to-Speech service then processes the selected text into a speech signal. In this current version of the Terabot project, no animated embodied agent (avatar) was employed. Instead, we decided to use a short film of a man wearing a hygiene mask. The man is slightly moving, and his eyes are blinking. Therefore, a patient sitting in front of the interface has the impression of having an online call with a therapist (see Figure 2).

Table 1. Statistics of training data, showing entries added by ChatGPT.

Intent	Training Data			ChatGPT-Added [%]
	Initial	ChatGPT	Total	
Greet	85	21	106	24.7
Goodbye	84	21	105	25.0
Affirm	245	38	283	15.5
Deny	39	11	50	28.2
Evade	80	49	129	61.3
Thank	14	30	44	214.3
Say_mood	202	412	614	204.0
Say_story	120	621	741	517.5
Choose_exercise	160	29	189	18.1
Say_adverbial	44	59	103	134.1
Say_avoid	14	46	60	328.6
Say_expload	19	44	63	231.6
Restart	22	0	22	0.0
Chitchat	198	329	527	166.2
Faq	199	0	199	0.0
Total	1525	1710	3235	112.1

Figure 2. Interface of the Terabot dialogue system.

A fragment of a therapeutic dialogue between the patient and Terabot is shown in Table 2 (emotion: fear). The patient is encouraged to answer several questions about the cause of their anxiety and their behavior when feeling it. After each response from the patient, its intent is recognized. Then, a following action is initiated, such as an affirmation, followed by another question, etc. Finally, after a successful conversation, the patient is encouraged to participate in a relaxation exercise.

Table 2. Sample Terabot–patient dialogue (fragment), with recognized intents (English translation).

Terabot	Patient
Do you recall a situation when you felt anxious?	
	Yes, uhm, I do very well. *(intent: Affirm)*
What caused your anxiety then?	
	I realized that I'd lost a fortune when, um, gambling. *(intent: Say_story)*
This is serious, indeed. Notice how quickly anxiety grows within you...	

To train the emotion-recognition module, initially we used CORTEX [18], a corpus of conversational emotional data for Polish, labeled with nine emotion classes consisting of 21k training sentences. However, its size was insufficient for our purposes. Therefore, this dataset was extended automatically by identifying semantically similar sentences in a web archive corpus called Common Crawl [1]. This approach resembled the semi-supervised learning (SSL) technique [37,38], in which a large unannotated dataset was assigned labels based on a classifier trained on a much smaller annotated dataset. We took the Polish subset of the Common Crawl archive, called hereinafter pCC , by filtering the whole set with the LanguageCrawl toolkit [39]. It resulted in a few billion web pages with some Polish content. As a next step, we selected only those web pages that contain at least 10 Polish sentences in a continuous manner. This led us to approx. 200 milion web pages, which we used ultimately for the CORTEX expansion. To measure semantic similarity between sentences, we used the sentence embeddings framework called Sentence-BERT (SBERT) [40]. This way, we computed embeddings of sentences coming from the CORTEX training dataset and pCC, and next, we evaluated their semantic similarity using cosine measure between embeddings. More precisely, for each sentence/prompt from the CORTEX training subset, we found the top 10 semantically closest results within pCC. In addition, we decided that their cosine similarity must be over a threshold of 0.8 (the value was set heuristically). This way, we expanded the CORTEX prompts with semantically similar candidate samples from pCC, and assigned them corresponding categories (mood, sentiment) of the given reference CORTEX prompt.

Next, we analyzed the expanded sentences with the assigned CORTEX categories using the RoBERTa classifier trained only on the original CORTEX training dataset. We predicted categories (mood, sentiment) for each "expanded" sentence. If they were the same as those assigned during the first expansion phase, we retained the given sentence in the extended dataset. If not, the sentence was removed. As a result, we created a new extended training dataset (named *CORTEX+pCC*), which is almost four times larger than the original CORTEX training dataset: it contained 79,400 sentences.

The DIET classifier, used for intent and slot recognition, was trained initially in 200 epochs by few-shot learning, i.e., using a limited amount of labeled data originating from mock therapeutic sessions. The dialogue system was developed using conversation-driven development, as suggested by the RASA developers. Initially, we sketched out some of the most likely dialogues in the *stories* file. Then, the dialogue system was tested by naïve users conversing with it. The model was retrained by adding misrecognized samples to the training data. The initial size of the training set was 1525 entries, as shown with details in Table 1. Frequent routines of intents and actions were added to the *rules* file. After several iterations, the dialogue system was exposed to patients.

2.2. Improving the Dialogue System Using ChatGPT

The initial informal trials with patients revealed that, despite positive experiences with healthy testers, many patients' utterances provoked wrong actions by Terabot. When analyzing system logs, it turned out that multiple intents were still misrecognized.

This demonstrated the need to expand the training dataset of the dialogue system, while ensuring it includes only relevant sentences and scenarios for patients. Continuing to develop the dialogue system by talking to patients would not be appropriate at this time. Since it is difficult to obtain both additional and meaningful sentences for the training database, we have decided to use ChatGPT to expand the data to train the DIET algorithm used in Terabot for recognizing intents.

When augmenting the training set using ChatGPT, we focused on the intents most often confused, e.g., *Say_story, Say_expload*. Various prompts were tried, as it turned out that the quality of the generated data depended strongly on how we formulated our questions for ChatGPT. For example, sometimes ChatGPT started to advise about coping with negative feelings instead of generating patients' reports on their negative feelings. However, we found heuristically that the most successful ways of generating training data were achieved when asking ChatGPT questions like:

- *Give a few examples with patients telling what recently made them anxious.*
- *Paraphrase briefly the sentence: "I usually avoid showing anger".*
- *Paraphrase the sentence: "I feel relaxed" using the feminine gender".*

The latter way of asking was motivated by the fact that without adding the comment on the feminine gender, ChatGPT mostly gave answers using the masculine gender (in Polish, gender impacts the form of verbs, adjectives, pronouns, etc.). Also, using the comment *"use alternately masculine or feminine gender"* was usually unsuccessful, as most of the resulting data were in the masculine gender.

Apart from the above-mentioned problem, the ChatGPT-generated data required further post-processing. We filtered out phrases that were not in Polish and those that were grammatically ill-formed in a very unlikely way, e.g., *"boi mi się"* (in English, literally: *"one fears me"*). Sample sentences that were generated with ChatGPT are shown in Appendix A. Both the examples in the target language (Polish) and their translations (for the reader's convenience) are presented. As a result of this data augmentation process, we more than doubled the size of the training data (see details in Table 1). We increased the training data for the *Say_story* intent by more than five times, and the training data for the *Say_avoid* intent by more than three times. Other training data were increased to a lesser extent. On the other hand, we did not enlarge the training data for *Restart* and *Faq* intents at all, as they were either perfectly recognized (the former) or rarely met during testing (the latter). In total, the resulting training set consisted of 3235 entries, compared to 1525 entries for the baseline model.

3. Experiments

Thirty-two adult psychiatric patients participated in therapeutic conversations with Terabot (19M, 13F, mean age 36.0 years, std. dev. 10.7 years). These patients were diagnosed with schizophrenia and were in a stable mental state. All patients were receiving pharmacological treatment, including antipsychotics, mostly in combination with other mood stabilizers or antidepressants. We informed them about the clinical trial, explained how it worked, and answered any questions they asked. They all signed a written consent and agreed to their voices being recorded.

Patients talked with Terabot for about 10–15 min in five sessions, one session daily. They could start with an exercise of their choice from the three available (*fear, anger, shame*), which could be repeated the next day. All conversations were recorded and then analyzed. We collected 2985 recordings in total. Out of them, we removed 183 recordings with unclear intent, leaving 2802 patients' utterances for further analyses. We calculated intent recognition accuracy as the ratio of correctly recognized intents against all intents. Based on 2802 of recordings, we compared results for the baseline model (i.e., trained without

the ChatGPT-generated data) and for the improved model, i.e., achieved with the dialogue model trained with the dataset enlarged using ChatGPT. We also compared the results of the patient tests with those of nine healthy subjects.

In addition to the above experiments, we also ran the intent recognition tests with clean data, i.e., assuming an ideal ASR. We are aware that such a scenario is not realistic; nevertheless, we wanted to verify what the impact of ASR errors on intent recognition actually was. We also performed statistical analyses using the Wilson score interval for a confidence level set to 95%.

As a supplementary analysis, we reviewed patient satisfaction surveys collected after the last therapeutic session. Every patient was asked to complete a satisfaction survey by answering on a Likert scale (i.e., 1 to 5) how much they agreed with the statements presented.

4. Results

The before-mentioned therapy sessions with patients took place between March and August 2023. During therapeutic sessions, especially in the first few minutes of talking, we noticed that most of the patients tended to talk quietly. However, one patient spoke at an unnaturally loud volume, practically shouting at the computer. Others, most likely due to side effects of medication, showed reduced articulation precision. We also observed one case of logorhea. Below, we present the results of our analyses conducted on 2802 recordings collected from 32 patients.

4.1. Intent Recognition Accuracy for the Model Enhanced with ChatGPT-Augmented Data

Figure 3 shows the results of intent recognition for the testers and for the patients using the baseline model, as well as for the patients using the model trained with data augmented with the ChatGPT-generated data. It shows that, unsurprisingly, for the vast majority of intents, the recognition accuracy decreased when moving from the healthy testers to the patients. The decrease was in a few cases severe, e.g., for *Say_story* it decreased from 85% to 31%, and for *Chitchat*, it decreased from 94% to 58%.

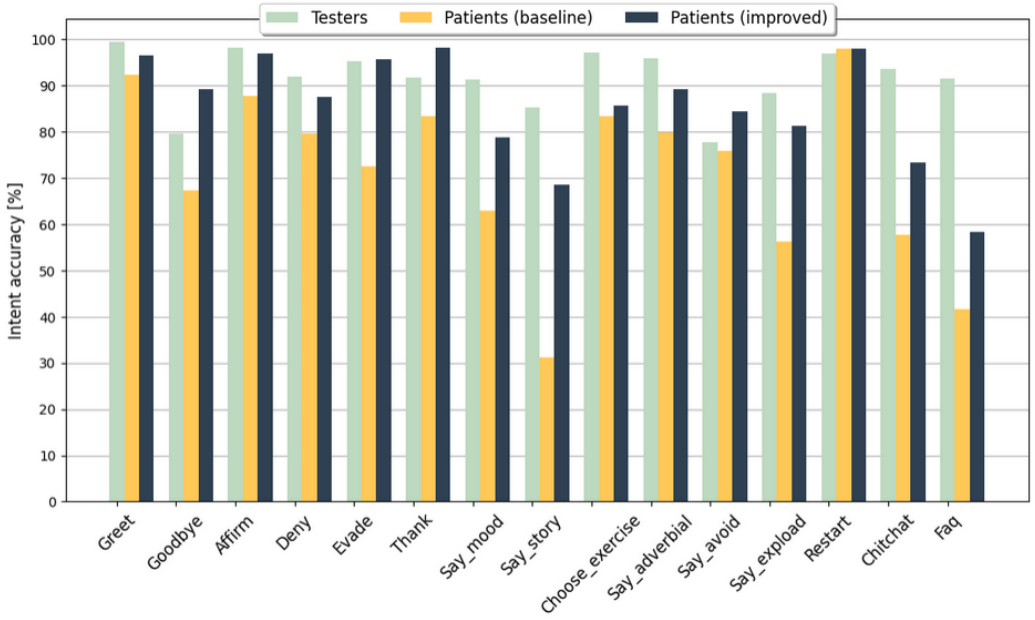

Figure 3. Intent recognition results for testers' speech, patients' speech with baseline model, and patients' speech with an improved model.

Luckily, the situation significantly improved after enhancing the model with the ChatGPT-generated data (see the third column for each intent in Figure 3). In all cases, the intent recognition accuracy increased, often approaching the value reached for testers

or even exceeding it, as in the cases of *Evade* or *Thank* intents. The most effective increase was observed for *Say_story* (from 31% to 69%), for *Say_expload* (from 56% to 81%) and for *Goodbye* (from 67% to 89%).

Interestingly, intent recognition for *Faq* also increased, even though no training data were added for that particular intent. Most likely, this is caused by the fact that by improving recognition for some intents (e.g., by adding training data for them), sometimes other intents are better recognized because they are less confused with the former ones. On average, augmenting the training set using ChatGPT allowed us to increase the intent recognition accuracy from 73.5% (weighted) to 86.6%, i.e., by 13% relatively. The error margins displayed in Table 3 confirm that this difference is statistically significant.

Table 3. Accuracy of intent recognition (in percentages) for system working on text with and without ASR errors. Error margins shown for confidence level 95%.

Input Data	DIET Intent Recognition Module	
	Baseline	**ChatGPT-Augmented**
Actual (with ASR errors)	73.52 ± 1.63	86.55 ± 1.26
Theoretical (for ideal ASR)	74.34 ± 1.62	86.62 ± 1.26

4.2. Intent Recognition Accuracy with Regard to ASR Errors

We wanted to refer intent recognition accuracy to the speech recognition accuracy yielded by the ASR module at entry to the Terabot system. Table 4 shows that WRR for patients yielded, on average, 91.4% (weighted), which was much lower than for the healthy testers (96.7%). Most likely, this was caused by decreased articulation precision, caused by antipsychotic medications, and, occasionally, a signal level that was too low. The intents *Say_story, Say_mood, Say_avoid* suffered most from low WRR. Surprisingly, the recognition was perfect for the *Restart_exercise* intent and even better than for the healthy testers (however, this intent was poorly represented in the patients' testset).

Table 4. Results of speech recognition for testers and patients. WRR and SRR are shown in percentages.

Intent	Testers			Patients		
	#Test	**WRR**	**SRR**	**#Test**	**WRR**	**SRR**
Greet	654	96.67	97.39	258	95.25	90.63
Goodbye	98	96.56	90.82	46	98.47	97.83
Affirm	556	97.40	94.63	671	94.14	90.69
Deny	74	100.00	100.00	88	87.62	90.91
Evade	207	99.62	98.53	69	92.44	92.75
Thank	12	100.00	100.00	54	94.23	87.50
Say_mood	159	98.38	94.12	244	89.10	79.84
Say_story	122	94.34	80.17	385	80.69	74.41
Choose_exer.	326	95.94	89.97	307	92.60	86.60
Say_adverb.	49	97.75	93.62	75	89.46	88.00
Say_avoid	9	100.00	100.00	58	87.35	86.21
Say_expload	26	90.81	70.37	16	95.78	81.25
Restart	65	95.63	93.18	253	95.69	94.05
Restart_exer.	22	98.68	95.46	3	100.00	100.00
Chitchat	269	96.98	85.04	262	92.72	83.94
Faq	131	89.41	68.75	13	97.69	92.31
Total	2779	–	–	2802	–	–
Weight.avg.	–	96.70	91.99	–	91.39	86.64

Undoubtedly, the imperfections of the ASR module impact the performance of intent recognition. One of the objectives of this study was to check the level of this impact. Table 3 compares the results of intent recognition accuracy for the actual system, i.e., with the ASR misrecognitions, with a theoretical one, i.e., with an error-free, ideal ASR.

When looking at the baseline model, we can observe that it would recognize intents with almost 74.3% accuracy on clean data, compared to about 73.5% on real data. So, it may indicate that the ASR was responsible for the accuracy decrease of about 1%. However, this difference lies within the error margin, so it would require a larger testset for verification. When looking at the model trained with the ChatGPT-augmented data, we can see that for clean data, the theoretical accuracy would be comparable with the baseline model, considering the error margin.

4.3. Satisfaction Survey

Figure 4 presents the results of the post-treatment questionnaire. Most patients highly rated the quality and speed of Terabot's speech, and highly appreciated the presence of another person during therapeutic sessions. The Likert scores for Terabot understanding patients' answers and emotions were 3.1 and 3.0, respectively. A total of 75% of the patients found the exercises proposed by the dialogue system helpful. What is very encouraging is that 13 patients confirmed they even liked talking to Terabot.

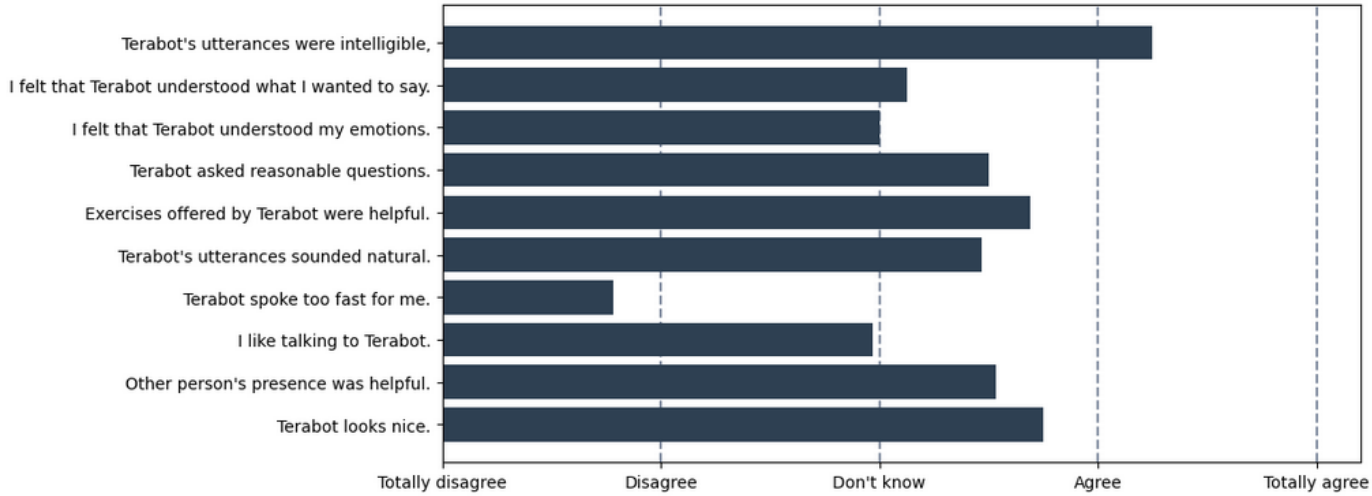

Figure 4. Patients' satisfaction on Likert scale.

5. Discussion

In this interdisciplinary research field, obtaining additional testers, patients, or sentences for the training database is demanding, and expanding the dialogue system's dataset is a major challenge. While conducting experiments with patients, it became clear that it would be very unlikely to provide the system with a large number of new training samples in a short time. We showed that using ChatGPT, and potentially other LLMs, is a way to quickly increase the size of the training dataset and improve the intent recognition accuracy. Here, the key benefit of our approach is the rapid and effective improvement. Compared to the 74% of intent recognition reported in [2], for the same set of intents, we have now achieved 86% accuracy.

A limitation of our solution is the fact of model hallucination. This is a well-known and important problem of LLMs, so researchers are attempting to find a solution. This is one of the reasons why we cannot use ChatGPT in a broader way than the one mentioned above; since we are dealing with a sensitive area, i.e., psychiatric therapy, we need to make sure that we can control what Terabot says.

Despite that, using ChatGPT in this specific manner allows for the efficient improvement of the training dataset. Expanding the dataset improves the dialogue model, leading to more accurate intent detection, so it would be possible to use Terabot more effectively in psychiatric therapy.

Looking at the results of the survey, we can summarize that most patients accepted our dialog system; however, not all of them. Both the Likert scores and the intent accu-

racy indicate that further work needs to be conducted to increase intent recognition and, consequently, patients' satisfaction.

6. Conclusions

In this paper, we described a successful case of employing one AI system (here, ChatGPT) to train another, a therapy-oriented spoken dialogue system named Terabot. We showed that by using properly formulated prompts to ChatGPT, we could quickly enlarge the training set for the intent recognition module of the dialogue system. In our case study, with 32 patients and 2802 test phrases, we improved intent recognition accuracy from 73.5% to 86.5%, which we found remarkable.

We also observed that most patients became seriously engaged in dialogue with Terabot, which is very encouraging. Twenty-four patients (out of 32) said that the relaxation exercises offered by our dialogue system helped them, and 13 even said they liked conversing with Terabot. However, the therapeutic effect of these exercises still needs to be verified.

We believe that our study nicely illustrates the potential of using generative language models to develop other AI-based tools. The greatest advantage of our approach was that the necessary data could be generated without involving end users (in this case, patients), which made it much faster, cheaper, and free from ethical constraints.

It is noteworthy that in our study, we used ChatGPT solely to generate data to train our system during its development. This means that we did not let ChatGPT control the conversation with the user, nor did we share users' data with it. We believe this approach can easily be followed in many other NLP-based applications where more training data are needed, yet it is necessary to retain a large level of control over the system's behavior and protect the privacy of users' data.

Despite the satisfactory results obtained in our study, we are aware of its limitations. Particular attention should be paid to the possibility of hallucinations in ChatGPT. Therefore, manual control of the results is still necessary. In the future, we plan to evaluate intent recognition accuracy on a larger group of patients and their conversations. We also plan to employ more advanced prompt engineering, which potentially can further improve the generation of training data.

Author Contributions: Conceptualization, K.G.-S., M.K. and A.J.; methodology, K.G.-S., I.S. and A.J.; software, R.R.; validation, K.G.-S., M.S. and A.J.; formal analysis, K.G.-S. and A.J.; investigation, K.G.-S. and A.J.; resources, R.R. and A.J.; data curation, K.G.-S. and R.R.; writing—original draft preparation, K.G.-S., M.S., I.S., M.K. and A.J.; writing—review and editing, K.G.-S. and A.J.; visualization, K.G.-S.; supervision, M.K. and A.J.; project administration, K.G.-S. and A.J.; funding acquisition, A.J. and M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Center for Priority Research Area Artificial Intelligence and Robotics of the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program. The APC was funded by a statutory activity subsidy from the Polish Ministry of Education and Science.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Institute of Psychiatry and Neurology in Warsaw, Poland, on 27 April 2022; resolution No. IV/2022.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The research data are available on request. For ethical reasons, the data are not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Example sentences provided to and received by ChatGPT.

Sample Answer to Terabot's Question in Polish *(with English Translation)*	**Example Sentences Paraphrased by ChatGPT in Polish *(with English Translation)***
Zazwyczaj unikam okazywania złości. *(I usually avoid showing anger.)*	Z reguły ukrywam swoją frustrację. *(I generally hide my frustration.)* Często maskuję swoje negatywne uczucia. *(I often mask my negative feelings.)* Zazwyczaj nie pokazuję swojej złości. *(I don't usually let my anger show.)* Na ogół unikam demonstracji mojego gniewu. *(In general, I avoid demonstrating my anger.)*
Zazwyczaj wybucham podczas okazywania złości. *(I usually burst out while showing anger.)*	Przeważnie tracę panowanie, gdy wybucham z powodu złości. *(Most of the time, I lose my temper when I explode out of anger.)* W większości przypadków moje reakcje na frustrację są gwałtowne. *(In most cases, my reactions to frustration are violent.)* Często moje reakcje na gniew są wybuchowe. *(A lot of times, my reactions to anger are explosive.)* Przeważnie tracę cierpliwość, gdy okazuję swoje negatywne uczucia. *(Mostly I lose patience when I show my negative feelings.)*
Moje ciało sztywnieje podczas odczuwania lęku, paraliżuje mnie. *(When I feel fear, my body stiffens up and paralyzes me.)*	W momencie odczuwania lęku, moje ciało staje się nieruchome, wywołując uczucie paraliżu. *(My body becomes motionless and feels paralyzed when I have anxiety.)* Pod wpływem lęku, moja fizyczna reakcja to zazwyczaj uczucie sztywności i paraliżu. *(Under anxiety, my physical reaction is usually a feeling of stiffness and paralysis.)* Kiedy odczuwam lęk, moje ciało jest jakby unieruchomione, pozostawiając mnie w stanie paraliżu. *(When I experience anxiety, my body is sort of immobilized, leaving me in a state of paralysis.)* W obliczu lęku, moje ciało wydaje się być sparaliżowane, i to uniemożliwia mi jakiekolwiek działanie. *(In the face of anxiety, my body seems paralyzed, preventing me from doing anything.)*

References

1. Kozłowski, M.; Gabor-Siatkowska, K.; Stefaniak, I.; Sowański, M.; Janicki, A. Enhanced Emotion and Sentiment Recognition for Empathetic Dialogue System Using Big Data and Deep Learning Methods. In Proceedings of the International Conference on Computational Science (ICCS 2023), Prague, Czech Republic, 3–5 July 2023; pp. 465–480. [CrossRef]
2. Gabor-Siatkowska, K.; Sowański, M.; Pudo, M.; Rzatkiewicz, R.; Stefaniak, I.; Kozłowski, M.; Janicki, A. Therapeutic Spoken Dialogue System in Clinical Settings: Initial Experiments. In Proceedings of the 30th International Conference on Systems, Signals and Image Processing, (IWSSIP 2023), Ohrid, North Macedonia, 27–29 June 2023; pp. 1–5. [CrossRef]
3. Rosenbaum, A.; Soltan, S.; Hamza, W.; Versley, Y.; Boese, M. LINGUIST: Language Model Instruction Tuning to Generate Annotated Utterances for Intent Classification and Slot Tagging. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; International Committee on Computational Linguistics: Gyeongju, Republic of Korea, 2022; pp. 218–241.
4. Chen, M.; Papangelis, A.; Tao, C.; Kim, S.; Rosenbaum, A.; Liu, Y.; Yu, Z.; Hakkani-Tur, D. PLACES: Prompting Language Models for Social Conversation Synthesis. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 814–838.
5. Zheng, C.; Sabour, S.; Wen, J.; Huang, M. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *arXiv* **2022**, arXiv:2202.13047.

6. Dino, F.; Zandie, R.; Abdollahi, H.; Schoeder, S.; Mahoor, M.H. Delivering Cognitive Behavioral Therapy Using A Conversational Social Robot. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019), Macau, China, 3–8 November 2019; pp. 2089–2095. [CrossRef]
7. Craig, T.K.; Rus-Calafell, M.; Ward, T.; Leff, J.P.; Huckvale, M.; Howarth, E.; Emsley, R.; Garety, P.A. AVATAR therapy for auditory verbal hallucinations in people with psychosis: A single-blind, randomised controlled trial. *Lancet Psychiatry* **2018**, *5*, 31–40. [CrossRef] [PubMed]
8. Stefaniak, I.; Sorokosz, K.; Janicki, A.; Wciórka, J. Therapy based on avatar-therapist synergy for patients with chronic auditory hallucinations: A pilot study. *Schizophr. Res.* **2019**, *211*, 115–117. [CrossRef]
9. Fernández-Caballero, A.; Navarro, E.; Fernández-Sotos, P.; González, P.; Ricarte, J.; Latorre, J.; Rodriguez-Jimenez, R. Human-Avatar Symbiosis for the Treatment of Auditory Verbal Hallucinations in Schizophrenia through Virtual/Augmented Reality and Brain-Computer Interfaces. *Front. Neuroinform.* **2017**, *11*, 64. [CrossRef] [PubMed]
10. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
11. Chen, Q.; Zhuo, Z.; Wang, W. BERT for joint intent classification and slot filling. *arXiv* **2019**, arXiv:1902.10909.
12. Zhong, V.; Xiong, C.; Socher, R. Global-Locally Self-Attentive Encoder for Dialogue State Tracking. In Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1458–1467. [CrossRef]
13. Su, P.H.; Gasic, M.; Mrkšić, N.; Barahona, L.M.R.; Ultes, S.; Vandyke, D.; Wen, T.H.; Young, S. On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems. In Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 2431–2441.
14. Sharma, S.; He, J.; Suleman, K.; Schulz, H.; Bachman, P. Natural Language Generation in Dialogue using Lexicalized and Delexicalized Data. In Proceedings of the International Conference on Learning Representations: Workshop Track, Toulon, France, 24–26 April 2017; pp. 1–6.
15. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* **2017**, *4*, e19. [CrossRef]
16. Spillane, B.; Saam, C.; Gilmartin, E.; Cowan, B.R.; Wade, V.P. ADELE: Evaluating and Benchmarking an Artificial Conversational Care Agent. In Proceedings of the Conversational Agents for Health and Wellbeing Workshop (CHI 2020), Honolulu, HI, USA, 25–30 April 2020.
17. Lugrin, B.; Pelachaud, C.; Traum, D., Eds. *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, 1st ed.; ACM: New York, NY, USA, 2022; Volume 48.
18. Zygadło, A.; Kozłowski, M.; Janicki, A. Text-Based emotion recognition in English and Polish for therapeutic chatbot. *Appl. Sci.* **2021**, *11*, 10146. [CrossRef]
19. Liu, X.; Zhou, G.; Kong, M.; Yin, Z.; Li, X.; Yin, L.; Zheng, W. Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method. *Systems* **2023**, *11*, 390. [CrossRef]
20. Liu, X.; Shi, T.; Zhou, G.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. Emotion classification for short texts: An improved multi-label method. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 306. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
22. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
23. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical Neural Story Generation. In Proceedings of the Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 889–898.
24. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019; pp. 1–6.
25. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, 3–10 March 2021; pp. 610–623.
26. Merrill, W.; Goldberg, Y.; Schwartz, R.; Smith, N.A. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1047–1060. [CrossRef]
27. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
28. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
29. Balas, M.; Ing, E.B. Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. *JFO Open Ophthalmol.* **2023**, *1*, 100005. [CrossRef]
30. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. *medRxiv* **2023**. [CrossRef]

31. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv* **2023**, arXiv:2302.04023.
32. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
33. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from Language Models. *arXiv* **2021**, arXiv:2112.04359.
34. Bunk, T.; Varshneya, D.; Vlasov, V.; Nichol, A. DIET: Lightweight Language Understanding for Dialogue Systems. *arXiv* **2020**, arXiv:2004.09936.
35. Jiao, A. An Intelligent Chatbot System Based on Entity Extraction Using RASA NLU and Neural Network. *J. Phys. Conf. Ser.* **2020**, *1487*, 012014. [CrossRef]
36. Vlasov, V.; Mosig, J.E.M.; Nichol, A. Dialogue Transformers. *arXiv* **2019**, arXiv:1910.00486.
37. Oliver, A.; Odena, A.; Raffel, C.A.; Cubuk, E.D.; Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.
38. Pudo, M.; Szczepanek, N.; Lukasiak, B.; Janicki, A. Semi-Supervised Learning with Limited Data for Automatic Speech Recognition. In Proceedings of the IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI 2022), Paris, France, 24–26 August 2022; pp. 136–141. [CrossRef]
39. Roziewski, S.; Kozłowski, M. LanguageCrawl: A generic tool for building language models upon common Crawl. *Lang. Resour. Eval.* **2021**, *55*, 1047–1075. [CrossRef]
40. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, 3–7 November 2019; pp. 3973–3983. [CrossRef]